p-ISSN: 2252-3006
e-ISSN: 2685-2411

# Teknik Data *Cleaning* Menggunakan Snowflake untuk Studi Kasus Objek Pariwisata di Bali

**Ni Putu Ayu Widiari, I Made Agus Dwi Suarjaya, Dwi Putra Githa**
Program Studi Teknologi Informasi, Fakultas Teknik, Universitas Udayana
Bukit Jimbaran, Bali, Indonesia, Telp. (0361) 701806
e-mail: ayuwidari@student.unud.ac.id, agussuarjaya@it.unud.ac.id, dwiputragitha@unud.ac.id

***Abstrak***

*Sejumlah besar data memiliki beberapa masalah yang sering ditemui seperti duplikasi data, ketidakkonsistenan data, dan ketidaklengkapan data. Variasi data yang dikumpulkan dari berbagai sumber akan mempengaruhi keakuratan hasil prediksi. Semakin banyak jumlah data yang dikumpulkan, pembersihan data manual hampir tidak mungkin karena memakan waktu dan rentan terhadap kesalahan. Untuk mempersingkat waktu dan mengurangi rentan kesalahan, diperlukan sebuah sistem yang dapat melakukan proses data cleaning secara otomatis. Tujuan dari proses data cleaning adalah menawarkan kualitas data yang lebih baik yang sangat membantu untuk memastikan data siap untuk tahap analisis. Salah satu tools pengolahan data yang dapat digunakan adalah Snowflake. Snowflake adalah tools pengolahan dengan basis query SQL yang dirancang untuk cloud. Data yang digunakan adalah tweet objek wisata di Bali melalui proses crawling data menggunakan Twitter API. Data yang dikumpulkan akan dibersihkan melalui dua tahap yaitu pembersihan Retweet dan kata noise, yang dilanjutkan dengan pembersihan untuk mencari tweet spesifik yang mengarah ke pariwisata Bali. Hasil proses cleaning objek wisata Bali pada 4 objek wisata yaitu Uluwatu, Sanur, Nusa Penida, dan Garuda Wisnu Kencana menunjukkan bahwa Nusa Penida merupakan objek wisata dengan jumlah penurunan yang signifikan dengan jumlah raw data yaitu 8087, cleaning tahap pertama yaitu 4770 data, dan cleaning tahap kedua adalah 2608 data.*

***Kata Kunci**: Data cleaning, Snowflake, Twitter, Bali*

***Abstract***

*Massive amounts of data have several problems that are often encountered such as data duplication, data inconsistency, and incomplete data. Data collected from the various sources will affect the accuracy of the prediction results. As more amount of data is collected, manual data cleaning is almost impossible because it is time consuming and prone to errors. To shorten time and reduce error prone, we need a system that can perform data cleaning processes automatically. The purpose of the data cleaning process is to offer better data quality which is very helpful to ensure the data is ready for the analysis phase. Cleaning data offers a better data quality which will be very helpful to ensure the data is ready for analysis. One of the data processing tools that can be used is Snowflake. Snowflake is a basis of SQL query tool that are designed for the cloud. The data used is tourist attractions tweets in Bali through data crawling process using the Twitter API. Data collected will be processed through two steps namely Retweet and noise word cleaning, and continued with specific tweets cleaning addressed to Bali. The result for 4 attractions in Bali namely* Uluwatu, Sanur, Nusa Penida*, and* Garuda Wisnu Kencana *shows that* Nusa Penida *is a tourist attraction with a significant amount decreased with the number of raw data that is 8087, first cleaning is 4770 data, and second cleaning is 2608 data.*

***Keywords**: Data cleaning, Snowflake, Twitter, Bali*

## 1. Pendahuluan

Data banyak digunakan untuk aktivitas organisasi dan mendorong keputusan bisnis. Data berkualitas buruk dapat berdampak negatif terhadap efektivitas dan efisiensi organisasi. Masalah kualitas data adalah salah satu kendala untuk menggunakan data secara efektif

karena data yang kotor dapat menyebabkan kesalahan keputusan [1]. Berbagai penelitian telah dilakukan selama bertahun-tahun untuk menemukan teknik pembersihan data terbaik untuk memecahkan masalah kualitas data.

Variasi data mungkin menjadi hambatan terbesar dalam penggunaan data untuk kepentingan analisis [2]. Jenis kesalahan yang berbeda seperti ketidaklengkapan, ketidakkonsistenan, duplikasi, dan nilai yang saling bertentangan terdapat dalam data yang akan mempengaruhi hasil analisis. Sejumlah penulis telah mengusulkan solusi untuk mengatasi masalah pembersihan data. Pembersihan data adalah operasi yang dilakukan pada data yang ada untuk menghilangkan anomali dan mendapatkan koleksi data[3]. Pembersihan data meliputi menghilangkan eror/kesalahan, menyelesaikan ketidakkonsistenan dan mentransformasikan data menjadi format yang seragam [4].

Proses pembersihan data adalah proses kompleks dan terdiri dari beberapa tahap yang meliputi penentuan aturan kualitas data, mendeteksi eror/kesalahan data, dan memperbaiki kesalahan. Pembersihan data dibagi menjadi pembersihan data tradisional dan pembersihan untuk data skala besar [5]. Metode pembersihan data tradisional disebut tradisional karena tidak cocok untuk menangani sejumlah data dengan volume dan skala besar. Potter's Wheel and Intelliclean adalah beberapa contoh pembersihan data secara tradisional [5]. Implementasi proses data *cleaning* banyak digunakan pada Sentimen Analisis, *text processing*, *Machine Learning*, dan *Natural Language Processing*.

Semakin banyak jumlah data yang dikumpulkan, pembersihan data manual hampir tidak mungkin karena memakan waktu dan rentan terhadap kesalahan. Untuk mempersingkat waktu dan mengurangi rentan kesalahan, diperlukan sebuah sistem yang dapat melakukan proses data *cleaning* secara otomatis. Salah satu *tools* yang digunakan untuk mengolah data dengan skala besar dan digunakan untuk proses data *cleaning* secara otomatis adalah Snowflake. Snowflake adalah data warehouse menggunakan mesin *database* SQL baru dengan arsitektur yang dirancang untuk *cloud*.

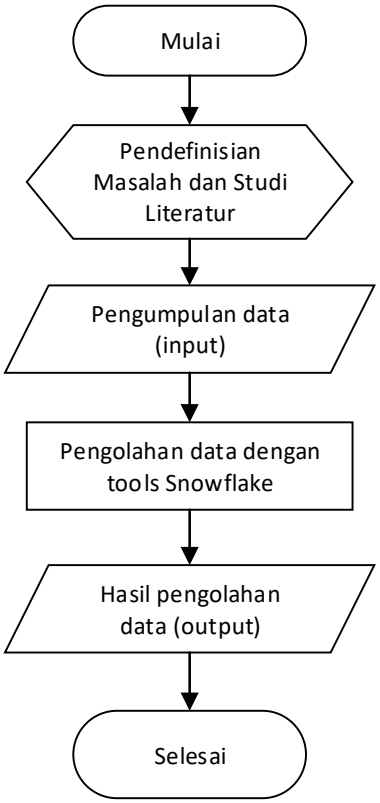
Penelitian terkait yang menggunakan proses data *cleaning* di dalamnya menggunakan proses *stopword removal* [6][7][8] dimana menghapus kata yang tidak memiliki makna yang termasuk pada *stopwords*. Penelitian lainnya [9] yang berjudul "*Text Based Approach For Similar Traffic Incident Detection from Twitter*" menggunakan kata kunci untuk proses *filtering*. Kata kunci ini diperoleh dengan mengamati tweet konten informasi lalu lintas. Sejumlah kata-kata penting yang sering muncul pada informasi trafik di korpus lalu dipilih sebagai kata kunci.

Penelitian ini mengusulkan *tools* data *cleaning* (*Big Data*) objek wisata di Bali menggunakan Snowflake. Proses *cleaning* meliputi proses menghilangkan kata *noise* yang dapat menganggu dalam proses analisis selanjutnya. Kata yang dihilangkan adalah kata kunci, username, hastags(#), email, Re*tweet*, ikon emosi, dan url.

## 2. Metodologi Penelitian

Penelitian terkait proses data cleaning menggunakan Snowflake dilakukan dalam empat Langkah yang dijabarkan melalui diagram alir dari tahapan penelitian seperti yang ditunjukkan pada Gambar 1 berikut.

p-ISSN: 2252-3006
e-ISSN: 2685-2411




Gambar 1. Diagram alir tahapan penelitian

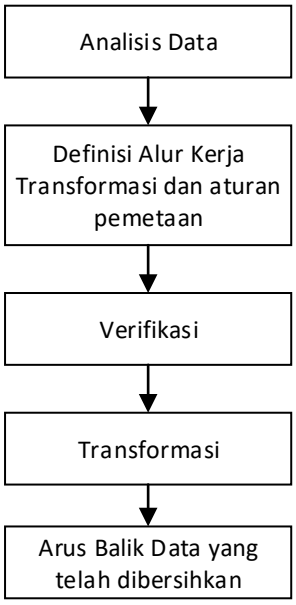
Penelitian ini dimulai dari mendefinisikan masalah yaitu bagaimana proses *data cleaning* menggunakan *tools* Snowflake, dilanjutkan dengan proses studi literatur mengenai Snowflake. Tahapan berikutnya adalah mengumpulkan data, data yang dikumpulkan berupa data hasil *crawling tweet*. *Crawling* merupakan tahapan pengumpulan data dengan melakukan pencarian melalui suatu kata kunci (*keywords*) tertentu dari media sosial. Data dikumpulkan dari Twitter menggunakan kata kunci terkait dan hastags(#) dengan objek wisata tiap kabupaten/kota di Bali secara berkala dari Oktober-Desember 2019 dengan mengakses Twitter API. Data *tweet* yang telah dikumpulkan diolah dengan *tools* Snowflake yang meliputi proses menghilangkan kata *noise* yang dapat menganggu dalam proses analisis selanjutnya. Kata yang dihilangkan adalah kata kunci, hastags(#),ikon emosi, simbol HTML*, username*, email, *Retweet*, dan url. Penelitian menghasilkan *outcome* berupa data *tweet* objek wisata bersih yang telah melalui proses *cleaning* menggunakan Snowflake.

## 3. Kajian Pustaka

### 3.1 *Data cleaning*

*Data cleaning* atau pembersihan data merupakan proses yang digunakan untuk mendeteksi, memperbaiki ataupun menghapus dataset, tabel, dan *database* yang korup atau tidak akurat. Istilah ini mengacu pada *dirty data* yang akan diganti, dimodifikasi atau dihapus setelah tahap identifikasi data yang tidak lengkap, tidak benar, tidak tepat, dan tidak relevan [10]. Proses *data cleaning* ini penting untuk mencegah data duplikat, membuat data lebih terstruktur, dan kompatibel [11]. Proses *data cleaning* terdiri dari 5 tahap yaitu; (1) Analisis Data, (2) Definisi alur kerja transformasi dan aturan pemetaan, (3) verifikasi, (4) Transformasi, (5) Arus balik data yang telah dibersihkan.

p-ISSN: 2252-3006
e-ISSN: 2685-2411




Gambar 2. Tahapan Proses *Data cleaning*

Langkah pertama dalam proses *Data cleaning* adalah menganalisis data untuk mengidentifikasi kesalahan dan ketidakkonsistenan yang terjadi di *database*. Dengan kata lain, fase ini disebut audit data di mana fase ini akan menemukan semua jenis anomali dalam *database*. Hasil dari langkah pertama adalah indikasi untuk setiap kemungkinan anomali apakah itu terjadi di dalam *database*. Selanjutnya, mendefinisikan alur kerja transformasi dan penghapusan anomali yang dilakukan mengikuti urutan operasi pada data. Hal ini ditentukan setelah analisis data untuk mendapatkan informasi tentang anomali yang ada. Jumlah langkah transformasi yang diperlukan tergantung pada jumlah sumber data, tingkat heterogenitas dan 'kekotoran' dari data. Untuk mengaktifkan pembuatan otomatis kode transformasi, transformasi terkait skema, dan langkah-langkah pembersihan harus ditentukan oleh kueri deklaratif dan bahasa pemetaan. Salah satu tantangan utama dalam hal ini fase adalah spesifikasi alur kerja dan aturan pemetaan yang akan diterapkan pada data kotor.

Langkah ketiga adalah tahap verifikasi. Dalam fase ini, kebenaran dan keefektifan transformasi alur kerja dievaluasi. Fase ini terdiri dari beberapa iterasi untuk memverifikasi semua kesalahan, Setelah data diverifikasi dan divalidasi, langkah-langkah transformasi akan dieksekusi. Proses transformasi membutuhkan sejumlah besar metadata seperti skema dan karakteristik data, pemetaan transformasi dan definisi alur kerja. Informasi terperinci tentang transformasi proses harus direkam untuk mendukung kualitas data. Setelah semua data kotor dibersihkan, maka data kotor seharusnya terganti dengan data bersih.

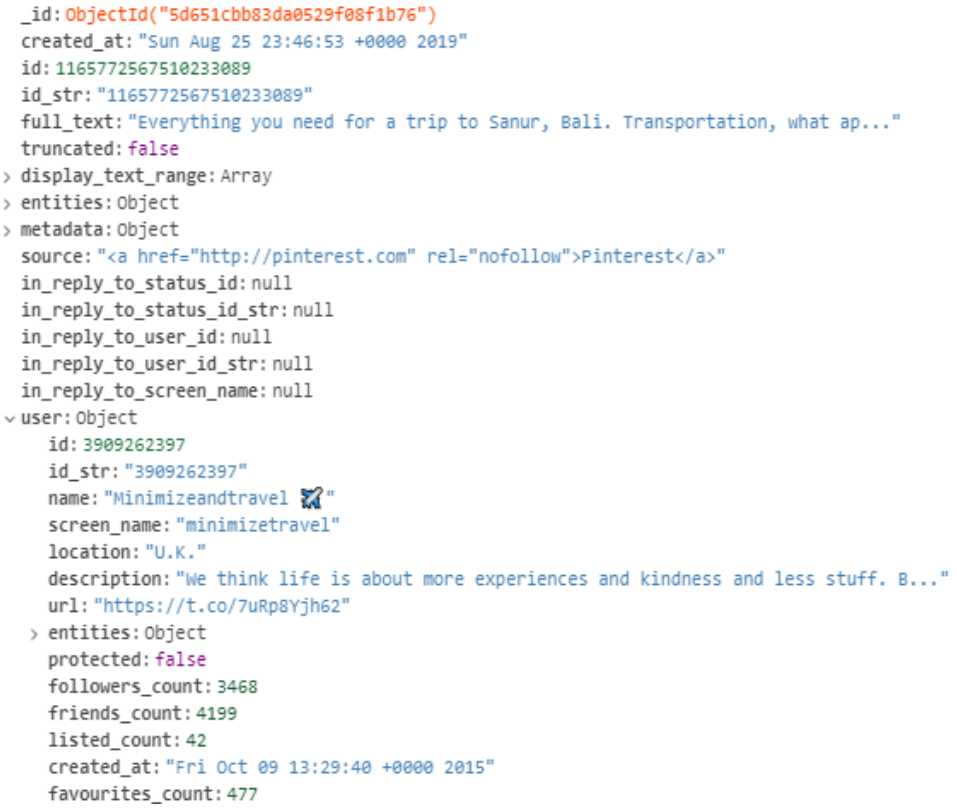
### 3.2 Snowflake

Snowflake adalah gudang data analitik yang disediakan sebagai Software-as-a-Service (SaaS). Snowflake menyediakan data warehouse yang lebih cepat dan lebih mudah digunakan, serta lebih fleksibel daripada penawaran data warehouse tradisional. Gudang data analitik Snowflake tidak dibangun di atas basis data yang ada atau platform perangkat lunak *big data* seperti Hadoop. Gudang data Snowflake menggunakan mesin *database* SQL baru dengan arsitektur unik yang dirancang untuk *cloud*. Untuk pengguna, Snowflake memiliki banyak kesamaan dengan gudang data perusahaan lainnya, tetapi juga memiliki fungsionalitas tambahan dan kemampuan unik.

## 4. Hasil dan Pembahasan

Berikut merupakan proses tahapan *cleaning* objek wisata di Bali. Proses awal adalah *crawling* data *tweet* dengan data hasil berupa data JSON. Proses dilanjutkan dengan proses *cleaning* menggunakan *query tools* Snowflake. Berikut merupakan hasil *crawling* data *tweet* objek wisata di Bali.

p-ISSN: 2252-3006
e-ISSN: 2685-2411




Gambar 3. *Tweet* JSON

Gambar 3 merupakan contoh *tweet* JSON hasil *crawling* data. Data *tweet* hasil *crawling* data merupakan data yang semi-struktur yang tidak terdiri dari *field* atau kolom dan tidak memiliki relasi seperti data pada *database* pada umumnya. *Raw tweet* data memiliki beberapa atribut seperti nama, ID, lokasi, deskripsi *tweet*, jumlah *follower*, jumlah teman, dan sebagainya.

```
CREATE TABLE table_name (id int, tanggal date, text varchar(1000),
retweet int, favorite int, matches string)
```

Kode Program 1. Sintaks *Create Table*

Kode program 1 merupakan sintaks membuat tabel dengan kolom id, tanggal, *text*, *retweet*, favorite, dan matches. Proses *cleaning tweet* objek wisata dengan Snowflake memerlukan tiga tabel (tb_*cleaning*, tb_filtering, tb_filter2) untuk menampung setiap data hasil *cleaning*. tb_*cleaning* merupakan tabel untuk menampung hasil penyamaan atribut kolom *tweet*. tb_filtering merupakan tabel untuk menyimpan hasil *cleaning* I yang merupakan penghapusan *tweet* yang mengandung kata sale, jual, dan *tweet* promosi. tb_filter2 merupakan tabel untuk menampung *tweet* bersih hasil *cleaning* II (penghapusan re*tweet*, mencari *tweet* yang spesifik mengarah ke objek wisata).

```
INSERT INTO tb_cleaning
select      c1:id::int      id,      :c1:created_at::date      tanggal,
c1:full_text::string   text,   c1:retweet_count::int   retweet_count,
c1:favorite_count::int favorite_count from tb_json
```

Kode Program 2. Sintaks *insert* tb_*cleaning*

p-ISSN: 2252-3006
e-ISSN: 2685-2411

Kode program 2 merupakan sintaks untuk menyisipkan data dari tabel sebelumnya ke tb_*cleaning* dengan hanya menyisipkan atribut yang diperlukan seperti id, tanggal, text, re*tweet*_count, dan favorite_count.

```
INSERT INTO tb_filtering
SELECT DISTINCT id, date, text, retweet_count, favorite_count,
lower(text) regexp
'\\b.*tanah\\b.*|^rt\\b.*|\\b.*jual\\b.*|\\b.*sale\\b.*|^available\\b
.*|\\b.*vacancy\\b.*|\\b.*Lowongan\\b.*|\\b.*sell\\b.*
|\\b.*lazada\\b.*|\\b.*voucher\\b.*|\\b.*villas\\b.*|\\b.*dog\\b.*|\\
b.*availabe\\b.*'
as matches from tb_cleaning where matches=false order by text
```

Kode Program 3. Sintaks insert tb_filtering

Kode program 3 merupakan sintaks untuk menyisipkan data dari tb_*cleaning* ke tb_filtering dimana data yang disisipkan adalah data yang tidak mengandung kata *noise*. *Query* SELECT DISTINCT merupakan *query* untuk tidak menampilkan data duplikat. *Query* lower(text) merupakan *query* untuk membuat seluruh karakter pada kolom text menjadi huruf kecil. *Query* REGEXP merupakan *query* untuk mencocokkan pola dalam data menggunakan karakter placeholder, yang disebut operator (\, . ,*). Karakter placeholder pada *query* diatas adalah tanah, jual, sale, lowongan, dan lain-lain yang merupakan kata *noise* yang harus dihilangkan untuk proses analisis selanjutnya.

| ID | DATE | TEXT | RETWEET_COUNT | FAVORITE_COUN | MATCHES |
|---|---|---|---|---|---|
| 1201537015529033... | 2019-12-02 | "Another great working view from I... | 0 | 1 | FALSE |
| 1195627897731657... | 2019-11-16 | "Bali, Uluwatu [OC] [3945×5917]" h... | 0 | 0 | FALSE |
| 1196161687155625... | 2019-11-17 | "Bali, Uluwatu. From u/flip14 on Red... | 0 | 0 | FALSE |
| 1199648373617418... | 2019-11-27 | "Don't ask what the world needs, a... | 0 | 0 | FALSE |
| 1182771608831905... | 2019-10-11 | "I was introduced to the dream of I... | 1 | 0 | FALSE |
| 1181947500380459... | 2019-10-09 | "I was introduced to the dream of I... | 2 | 1 | FALSE |
| 1181237825305501... | 2019-10-07 | "I was introduced to the dream of I... | 0 | 0 | FALSE |

Gambar 4. Detail hasil tb_filtering

Gambar 4 merupakan detail hasil tb_filtering yang terdiri dari 6 kolom yaitu kolom id, date, text, retweet_count, favorite_count, dan matches. Hasil *filtering* adalah mencari kumpulan *tweet* yang tidak mengandung kata *noise* seperti pada kode program 3.

```
INSERT INTO tb_filter2
SELECT * FROM tb_filtering where text not LIKE 'RT%'
```

Kode Program 4. Sintaks insert tb_filter2

Kode Program 4 merupakan sintaks untuk menyisipkan data dari tb_filtering ke tb_filter2 dimana data tidak mengandung kata ‘RT’ pada kolom text. *Query* SELECT * FROM merupakan *query* untuk menampilkan seluruh data dari tb_filtering. *Query* LIKE merupakan *query* untuk mencari pola tertentu dalam kolom pada suatu tabel.

p-ISSN: 2252-3006
e-ISSN: 2685-2411

| ID | DATE | TEXT | RETWEET | FAVORITE | MATCHES |
|---|---|---|---|---|---|
| 12031808198142402... | 2019-12-07 | 7 Des \| Dibutuhkan : 1 (satu) Cook 1 (satu... | 0 | 0 | false |
| 11922293881605242... | 2019-11-06 | 7 November \| Triangle English Centre me... | 0 | 0 | false |
| 11891938133544345... | 2019-10-29 | 7 Tips Menyaksikan Kecak Fire Dance Ul... | 0 | 0 | false |
| 1192563139713228801 | 2019-11-07 | 8 Nov \| Dibutuhkan segera karyawan di b... | 0 | 0 | false |
| 11868947636491673... | 2019-10-23 | 8,5 years ago when you were witnessed ... | 0 | 0 | false |
| 1185184985017307137 | 2019-10-18 | @515SunLight เอาเป็นสถานที่ ไม่เอาเป็นเมือง... | 0 | 1 | false |
| 12110548688505282... | 2019-12-28 | @56Hz I have a soft spot for Uluwatu. | 0 | 0 | false |
| 1203492414154887170 | 2019-12-08 | @ADNInfoDrive Uluwatu Beach, Bali ADN... | 2 | 0 | false |
| 12076073597170032... | 2019-12-19 | @A_DeyJKT48 Udeh lo kaga cocok pake ... | 0 | 0 | false |
| 12057204275001917... | 2019-12-14 | @AdjieSanPutro @arievrahman @whatra... | 0 | 0 | false |

Gambar 5. Detail hasil tb_filter2

Gambar 5 merupakan detail hasil tb_filter2 yang terdiri dari 6 kolom yaitu kolom id, date, text, retweet_count, favorite_count, dan matches. Hasil *filtering* adalah mencari kumpulan *tweet* yang tidak mengandung kata RT atau *Retweet* seperti pada kode program 4.

```
SELECT * FROM tb_filter2 where text LIKE '%pura%' OR text LIKE
'%uluwatu%' OR text LIKE '%pantai%' OR text LIKE '%beach%'
OR text LIKE '%bali%' OR text LIKE '%#uluwatu%' OR text LIKE
'%travel%' OR text LIKE '%holiday%'
```

Kode Program 5. Sintaks select tb_filter2

Kode program 5 merupakan sintaks untuk menampilkan data dari tb_filter2 dimana data yang ditampilkan adalah data yang memiliki pola kata (pura, uluwatu, pantai, beach, bali, uluwatu, travel, holiday) pada kolom text.

| ID | DATE | TEXT | RETWEET | FAVORITE | MATCHES |
|---|---|---|---|---|---|
| 1196161687155625988 | 2019-11-17 | "Bali, Uluwatu. From u/flip14 on Reddit #u... | 0 | 0 | false |
| 1199648373617418240 | 2019-11-27 | "Don't ask what the world needs, ask wh... | 0 | 0 | false |
| 1192743279172829191 | 2019-11-08 | "Pura Luhur Uluwatu adalah salah satu d... | 3 | 1 | false |
| 1207930218410725377 | 2019-12-20 | "maaf ya sayang kacamatanya potek lagi... | 0 | 0 | false |
| 1195068139224076290 | 2019-11-14 | #Bali #Indonesia is one of the most amaz... | 0 | 2 | false |
| 1181805762344964096 | 2019-10-09 | #Bali is an #Indonesian #island known for... | 0 | 1 | false |
| 1195639542574960640 | 2019-11-16 | #Bali, Uluwatu [OC] [3945×5917] #travel ... | 0 | 0 | false |

Gambar 6. Hasil *filtering* tweet spesifik

Gambar 6 merupakan detail hasil tb_filter2 yang terdiri dari 6 kolom yaitu kolom id, date, text, retweet_count, favorite_count, dan matches. Hasil *filtering* adalah mencari kumpulan *tweet* yang spesifik dan mengandung pola kata objek wisata (pura, uluwatu, pantai, beach, bali, uluwatu, travel, holiday)

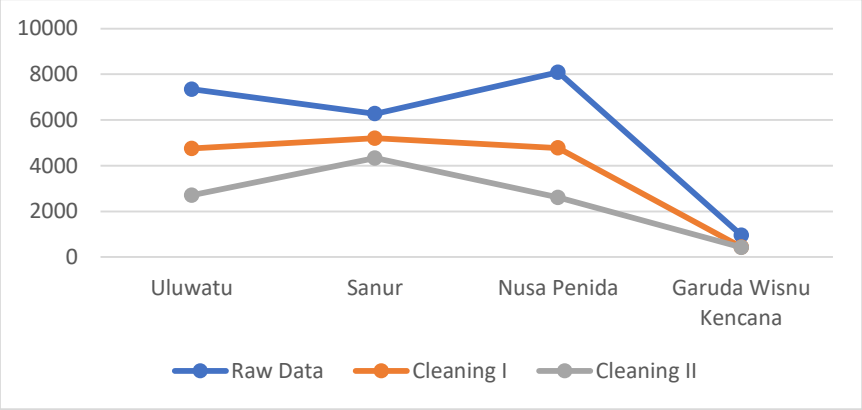
Tabel 1. Hasil Proses *Cleaning* Objek Wisata dengan Snowflake

| No | Objek Wisata | Raw Data | *Cleaning* I | *Cleaning* II |
|---|---|---|---|---|
| 1 | Uluwatu | 7346 | 4754 | 2703 |
| 2 | Sanur | 6269 | 5198 | 4330 |
| 3 | Nusa Penida | 8087 | 4770 | 2608 |

p-ISSN: 2252-3006
e-ISSN: 2685-2411

| No | Objek Wisata | Raw Data | *Cleaning* I | *Cleaning* II |
|---|---|---|---|---|
| 4 | Garuda Wisnu Kencana | 957 | 439 | 421 |

Empat objek wisata ini digunakan karena adanya limitasi dari penggunaan *tools* Snowflake. *Tools* Snowflake untuk versi trial terdapat limitasi untuk *database* penyimpanan. Tabel 1 merupakan hasil proses *cleaning* objek wisata di Bali dengan jumlah 4 objek wisata yaitu Uluwatu, Sanur, Nusa Penida, dan Garuda Wisnu Kencana. Hasil dari proses *cleaning* menunjukkan adanya penurunan siginifikan jumlah *tweet* dari *raw* data. Penurunan proses *cleaning* pertama disebabkan adanya *retweet* yang mengandung *spam*, *tweet* promosi, dan penjulan. Sedangkan penurunan yang terjadi pada proses *cleaning* tahap kedua disebabkan karena pencarian *tweet* spesifik yang mengarah ke pariwisata Bali dengan menggunakan kata kunci (contoh: uluwatu bali, #sanurbali). Nusa Penida merupakan objek wisata dengan jumlah penurunan yang signifikan dengan jumlah *raw data* yaitu 8087, *cleaning* I yaitu 4770 data, dan *cleaning* II adalah 2608 data dengan *tools* Snowflake.




Gambar 7. Grafik Hasil *Cleaning* Objek Wisata di Bali

Pada gambar 7, grafik Garuda Wisnu Kencana merupakan objek wisata dengan raw data terendah yaitu 957 data, hasil proses *cleaning* I yaitu 439 data, dan hasil proses *cleaning* II yaitu 421 data. Penurunan terhadap *tweet* setelah proses *cleaning* memiliki makna bahwa *raw* data yang dikoleksi selama periode Oktober-Desember 2019 sebagian besar merupakan *tweet* hasil *Retweet*, *tweet* berupa *spam, tweet* penjualan, dan promosi.

## 5. Kesimpulan

*Data cleaning* atau pembersihan data merupakan proses yang digunakan untuk mendeteksi, memperbaiki ataupun menghapus dataset, tabel, dan *database* yang korup atau tidak akurat. Tahapan dari proses *data cleaning* yang terdapat dalam penelitian ini adalah data hasil *crawling* yang disimpan akan dibersihkan melalui 2 proses pembersihan yaitu pembersihan re*tweet* dan kata *noise* seperti sale, jual, dan *tweet* promosi, dan dilanjutkan dengan pembersihan untuk mencari *tweet* spesifik yang mengarah ke pariwisata Bali. Hasil proses *cleaning* objek wisata Bali pada 4 objek wisata yaitu Uluwatu, Sanur, Nusa Penida, dan Garuda Wisnu Kencana menunjukkan bahwa Nusa Penida merupakan objek wisata dengan jumlah penurunan yang signifikan dengan jumlah *raw data* yaitu 8087, *cleaning* I yaitu 4770 data, dan *cleaning* II adalah 2608 data. Penurunan terhadap *tweet* setelah proses *cleaning* memiliki makna bahwa *raw* data yang dikoleksi selama periode Oktober-Desember 2019 sebagian besar merupakan *tweet* hasil *Retweet*, *tweet* berupa *spam, tweet* penjualan, dan promosi.

## Daftar Pustaka

[1] O. Azeroual, G. Saake, and M. Abuosba, “Azeroual_jdimv16i1_2,” vol. 16, no. 1, 2018.
[2] W. Swapnil and Y. Anil, “Big Data: Characteristics, Challenges and Data Mining,” *Int. J. Comput. Appl.*, pp. 975–8887, 2016.

[3] S. Devi and A. Kalia, “Study of Data Cleaning & Comparison of Data Cleaning Tools,” *Int. J. Comput. Sci. Mob. Comput.*, vol. 4, no. 3, pp. 360–370, 2015.
[4] H. Woo *et al.*, “Application of efficient data cleaning using text clustering for semistructured medical reports to large-scale stool examination reports: Methodology study,” *J. Med. Internet Res.*, vol. 21, no. 1, 2019, doi: 10.2196/10013.
[5] F. Ridzuan and W. M. N. Wan Zainon, “A review on data cleansing methods for big data,” *Procedia Comput. Sci.*, vol. 161, pp. 731–738, 2019, doi: 10.1016/j.procs.2019.11.177.
[6] I. Made Suwija Putra, N. Putu Ayu Widiari, and I. Wayan Gunaya, “Implementasi Generalized Vector Space Model (GVSM) dalam Pencarian Buku di Perpustakaan,” *J. Ilm. Merpati (Menara Penelit. Akad. Teknol. Informasi)*, vol. 7, no. 1, p. 86, 2019, doi: 10.24843/jim.2019.v07.i01.p10.
[7] N. K. Widyasanti, I. K. G. Darma Putra, and N. K. Dwi Rusjayanthi, “Seleksi Fitur Bobot Kata dengan Metode TFIDF untuk Ringkasan Bahasa Indonesia,” *J. Ilm. Merpati (Menara Penelit. Akad. Teknol. Informasi)*, vol. 6, no. 2, p. 119, 2018, doi: 10.24843/jim.2018.v06.i02.p06.
[8] A. R. Chrismanto and Y. Lukito, “Identifikasi Komentar Spam Pada Instagram,” *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 8, no. 3, p. 219, 2017, doi: 10.24843/lkjiti.2017.v08.i03.p08.
[9] M. Ermawati and J. L. Buliali, “Text Based Approach For Similar Traffic Incident Detection from Twitter,” *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 9, no. 2, p. 63, 2018, doi: 10.24843/lkjiti.2018.v09.i02.p01.
[10] R. R. Rerung, “Penerapan Data Mining dengan Memanfaatkan Metode Association Rule untuk Promosi Produk,” *J. Teknol. Rekayasa*, vol. 3, no. 1, p. 89, 2018, doi: 10.31544/jtera.v3.i1.2018.89-98.
[11] B. Cohen *et al.*, “Challenges Associated With Using Large Data Sets for Quality Assessment and Research in Clinical Settings,” *Policy, Polit. Nurs. Pract.*, vol. 16, no. 3–4, pp. 117–124, 2015, doi: 10.1177/1527154415603358.