Klasifikasi *Dangerous Speech* pada Media Sosial Twitter Menggunakan *Unsupervised dan Supervised Learning*

****

**Disusun Oleh:**

**Kelompok 05**

1. Muh. Nur Fajrin Amiruddin (05111940000005)
2. Muhammad Arif Faizin (05111940000060)
3. Kelvin Andersen (05111940000080)
4. Ihsannur Rahman Qalbi (05111940000090)

**DEPARTEMEN INFORMATIKA**

**FAKULTAS TEKNOLOGI ELEKTRO DAN INFORMATIKA CERDAS**

**SURABAYA**

**2022**

# DAFTAR ISI

[**DAFTAR ISI**](#_rg8k02llfh89) **2**

[**ABSTRAK**](#_9rt5t78g1lsu) **3**

[**I. PENDAHULUAN**](#_mg09srsyss2n) **4**

[1.1 Latar Belakang](#_jkr0wbdlszcg) 4

[1.2 Rumusan Masalah](#_dm0mbqejcwaa) 4

[1.3 Tujuan Penelitian](#_nlfq1ogn6ksu) 4

[1.4 Manfaat dan Dampak Keberlanjutan Penelitian](#_oywmsruef540) 4

[**II. METODE PELAKSANAAN**](#_46fppy7ivwpr) **5**

[2.1 Studi Literatur](#_q0j5gmds1j5) 5

[2.1.1 Web Scraping](#_u6bob0kq3yfg) 5

[2.1.2 Stopword](#_3p1p78kyg5kg) 5

[2.1.3 Tokenisasi](#_xn2et2v89xhs) 5

[2.1.4 K-Means Clustering](#_x464keqca0ga) 5

[2.1.5 Word2Vec](#_hkauc1u7zaqe) 5

[2.1.6 TF-IDF (Term Frequency Inverse Document Frequency)](#_if6ja7w6uri3) 6

[2.1.7 Cluster Analysis](#_4bualpo6cbl1) 6

[2.2 Analisis Perbandingan Penelitian](#_olrmwr73eyfz) 6

[2.2 Langkah Pengerjaan](#_4sckdvpxkrf8) 6

[2.3 Pengumpulan Data](#_5g5ctako7l0m) 6

[2.4 Pre-processing](#_e4hq6tb7wima) 8

[2.5 Eksplorasi Data](#_uixf8iguh3gk) 8

[2.6 Ekstraksi Fitur](#_bqacilurwm4l) 10

[2.7 Metode Clustering](#_loy8271ro3tl) 10

[**III. HASIL DAN PEMBAHASAN**](#_3g1ixuwvcncr) **11**

[3.1 Word2vec](#_xx9eyowz2ftw) 11

[3.2 TF-IDF](#_x1ydbfbl7jn5) 12

[3.3 CountVectorizer](#_4rj3bk6di1kc) 14

[**IV. KESIMPULAN**](#_qwscupbhv3dm) **18**

[**DAFTAR PUSTAKA**](#_5xnp4c1jm4q6) **19**

# 

# ABSTRAK

**Twitter** adalah sebuah sosial media yang dimana seluruh **user** bebas berpendapat dan menulis apapun yang mereka inginkan. Karena kebebasan ini, tidak jarang muncul tulisan - tulisan dari yang kurang berkenan dalam bentuk **hate speech** dan **dangerous speech**. Maka dari itu, *clustering* dilakukan untuk membantu mengerti apa yang dibahas oleh suatu akun *Twitter* yang menyerang oposisi. Dalam kasus ini akun tersebut adalah “@\_\_Bungs”. Hasil *clustering* yang bagusdidapatkan menggunakan fitur ekstraksi Word2vec dan CountVectorizer dimana *silhouette score* tertinggi yang didapat adalah 0.4618 menggunakan metode fitur ekstraksi Word2vec dan metode *clustering* K-Means yang menghasilkan dua *cluster*. Namun, metode ini tidak dapat digunakan untuk melakukan klasifikasi *dangerous speech*. Metode feature extraction yang dapat digunakan untuk mengklasifikasi tweet yang dapat mengandung dangerous speech adalah CountVectorizer. Silhouette score pada metode ini untuk jumlah cluster 5 adalah 0.017 dan terdapat satu cluster yang mengandung tweet yang dapat mengandung dangerous speech.

Kata Kunci: **Twitter,** ***clustering*, hate speech, dangerous speech**

# I. PENDAHULUAN

## 1.1 Latar Belakang

Di era modern ini twitter termasuk menjadi salah satu media sosial yang sering digunakan orang banyak karena twitter menyediakan tempat untuk penggunanya mengunggah tulisan dan berinteraksi dengan pengguna lainnya dengan pembatasan/peraturan yang minim.

Karena kebebasan yang diberikan kepada penggunanya, terkadang muncul beberapa unggahan yang kurang berkenan untuk dikonsumsi berbagai kalangan masyarakat seperti *hate speech* dan *dangerous speech.*

Ada banyak faktor yang mendorong individu untuk menghasilkan *hate speech*, faktor faktor ini dapat berupa permasalahan emosi yang personal, *hoax*, dan bahkan untuk hiburan individu itu sendiri (Beryandhi, 2020). Jenis *hate speech* yang muncul di media sosial sangat beragam, dapat berupa SARA, *body shaming*, bahkan sampai dorongan untuk bunuh diri

Riset ini bertujuan untuk menentukan topik *hate speech* atau *dangerous speech* yang dimaksud pada suatu tulisan yang diunggah oleh seorang atau beberapa pengguna twitter. Untuk mencapai kesimpulan topik apa yang sedang dibahas pengguna, diperlukan beberapa langkah yang harus dilakukan terlebih dahulu, dari pengumpulan data, klasifikasi dan kemudian analisis.

## 1.2 Rumusan Masalah

Berdasarkan latar belakang diatas, rumusan masalah untuk karya tulis ilmiah ini adalah sebagai berikut

1. Bagaimana proses pengumpulan dan persiapan data *Tweet* dari Twitter?
2. Bagaimana melakukan *topic modelling* pada *Tweet*?

## 1.3 Tujuan Penelitian

Target yang ingin dicapai dalam karya ilmiah ini adalah berupa kesimpulan yang menunjukan topik apa yang sedang dibahas seseorang atau sekelompok pengguna twitter dalam unggahannya.

## 1.4 Manfaat dan Dampak Keberlanjutan Penelitian

Dengan mendapatkan topik *hate speech* yang sedang dibahas, kita dapat menentukan relasi munculnya *hate speech* dengan event-event dengan dampak tertentu.

# II. METODE PELAKSANAAN

## 2.1 Studi Literatur

Studi literatur melalui artikel ilmiah dilaksanakan untuk mendapatkan konsep dasar penelitian. Hal tersebut meliputi *web scraping*, *stopword*, tokenisasi, *K-means clustering*, *word2vec*, TF-IDF. Kemudian peneliti juga menggali tentang *cluster analysis*. Hasil studi literatur yang diperoleh adalah sebagai berikut

### 2.1.1 *Web Scraping*

*Web scraping* adalah teknik mengekstrak data yang tidak terstruktur dari website, dan mengubahnya ke data terstruktur yang bisa disetorkan dan dianalisa dalam database. *Web scraping* juga dapat disebut dengan *web data extraction*, *web harvesting* atau *screen scraping*. *Web scraping* adalah salah satu bentuk dari *data mining*. Tujuan utama dari *web scraping* adalah untuk mengekstrak informasi dari sebuah website dan mengubahnya ke struktur yang dapat dipahami seperti *spreadsheet*, *database*, atau *comma-separated values* (CSV) file. (Sirisuriya, 2015)

### 2.1.2 *Stopword*

*Stopword* merupakan kata-kata umum yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki arti. *Stopword* umumnya dimanfaatkan dalam task *information retrieval*. Contoh *stopword* untuk bahasa Inggris diantaranya “of”, “the”. Sedangkan untuk bahasa Indonesia diantaranya “yang”, “di”, “ke” (Assuja, 2016).

### 2.1.3 Tokenisasi

Tokenisasi dilakukan untuk memenggal tiap kata atau *term* yang terpisah oleh anda baca seperti spasi, titik, koma, tanda tanya, dsb. *Term-term* atau token-token tersebut dipilih kembali sebagai kata yang paling mewakili tiap label sentimen pada tahap ekstraksi fitur untuk fase *training* (Assuja, 2016).

### 2.1.4 K-Means Clustering

*K-Means* adalah teknik yang cukup sederhana dan cepat dalam proses *clustering* objek (*clustering*) serta mampu mengelompokkan data dalam jumlah cukup besar. Berikut adalah algoritma dari metode *K*-*Means*.

1. Memilih secara acak k *centroid* awal dalam data
2. Menentukan jarak setiap kata terhadap pusat *cluster*
3. Mengelompokkan setiap kata berdasarkan kedekatannya dengan *centroid* (jarak terkecil) menggunakan *Euclidean distance*.

### 2.1.5 Word2Vec

Word2Vec merupakan algoritma representasi vektor kata yang mampu mencapai kinerja terbaik dalam NLP (*Natural Language Processing*) dengan cara mengelompokan kata yang serupa memiliki vektor yang sama. Word2Vec menghitung representasi kata ke dalam vektor menggunakan *neural network*. Vektor kata yang dihasilkan adalah vektor ruang dimensi yang menangkap makna semantik dari kata.

### 2.1.6 TF-IDF (Term Frequency Inverse Document Frequency)

*Term Frequency-Inverse Document Frequency* (TF-IDF) digunakan untuk mengukur seberapa penting suatu kata dalam dokumen. Frekuensi kemunculan setiap kata dalam dokumen diberikan sebagai skema pembobotan kata yang diasumsikan bahwa *stopwords* yang paling sering muncul dalam teks dihapus terlebih dahulu. Kata-kata yang diberikan bobot (*weight*) dihitung frekuensi kemunculannya dalam sebuah dokumen.

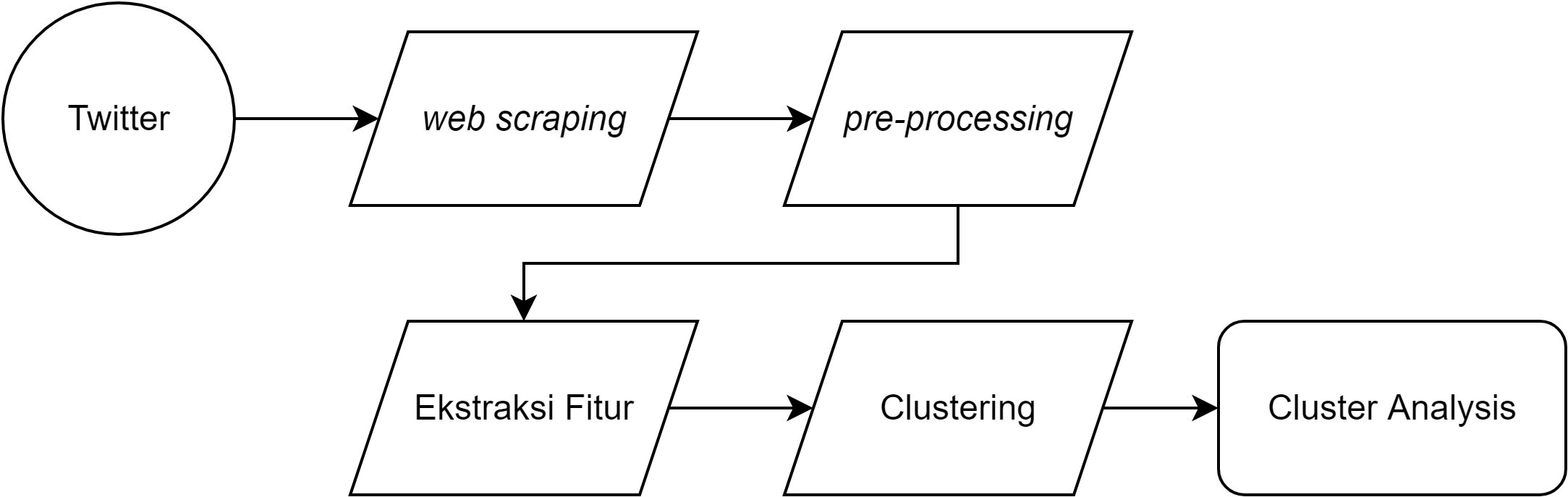
### 2.1.7 Cluster Analysis

*Cluster analysis* adalah teknik multivariat yang dapat bersifat objektif dalam menilai dan mengklasifikasikan objek berdasarkan karakteristiknya. Sehingga, tujuan dari pengelompokan ini adalah menggabungkan objek yang serupa (atau mirip) satu sama lain dalam satu grup dan memisahkannya atau membedakannya dengan objek di luar grup (Muthahharah, 2021).

## 2.2 Analisis Perbandingan Penelitian

## 2.2 Langkah Pengerjaan

Dalam melakukan penelitian ini, terdapat beberapa tahapan yang dapat dilakukan untuk mendapatkan hasil analisis dari dataset. Diantaranya terdapat pada Gambar 2.1 berikut

[](https://app.diagrams.net/?page-id=HXKbnfuub7f8G5PgmI1q&scale=auto#G1qLIsltYGSbXLyW8bzNjxvnJldMOJTBId)

**Gambar 2.1** Diagram langkah pengerjaan

Proses *web scraping* dilakukan untuk pengumpulan data mentah, dilanjut dengan *pre-processing* untuk mendapatkan data yang dapat diolah selanjutnya. Kemudian dilanjutkan dengan proses ekstraksi fitur, dimana proses ini sangat penting untuk menentukan kapabilitas data. Selanjutnya, untuk melakukan *cluster analysis* diperlukan *clustering*.

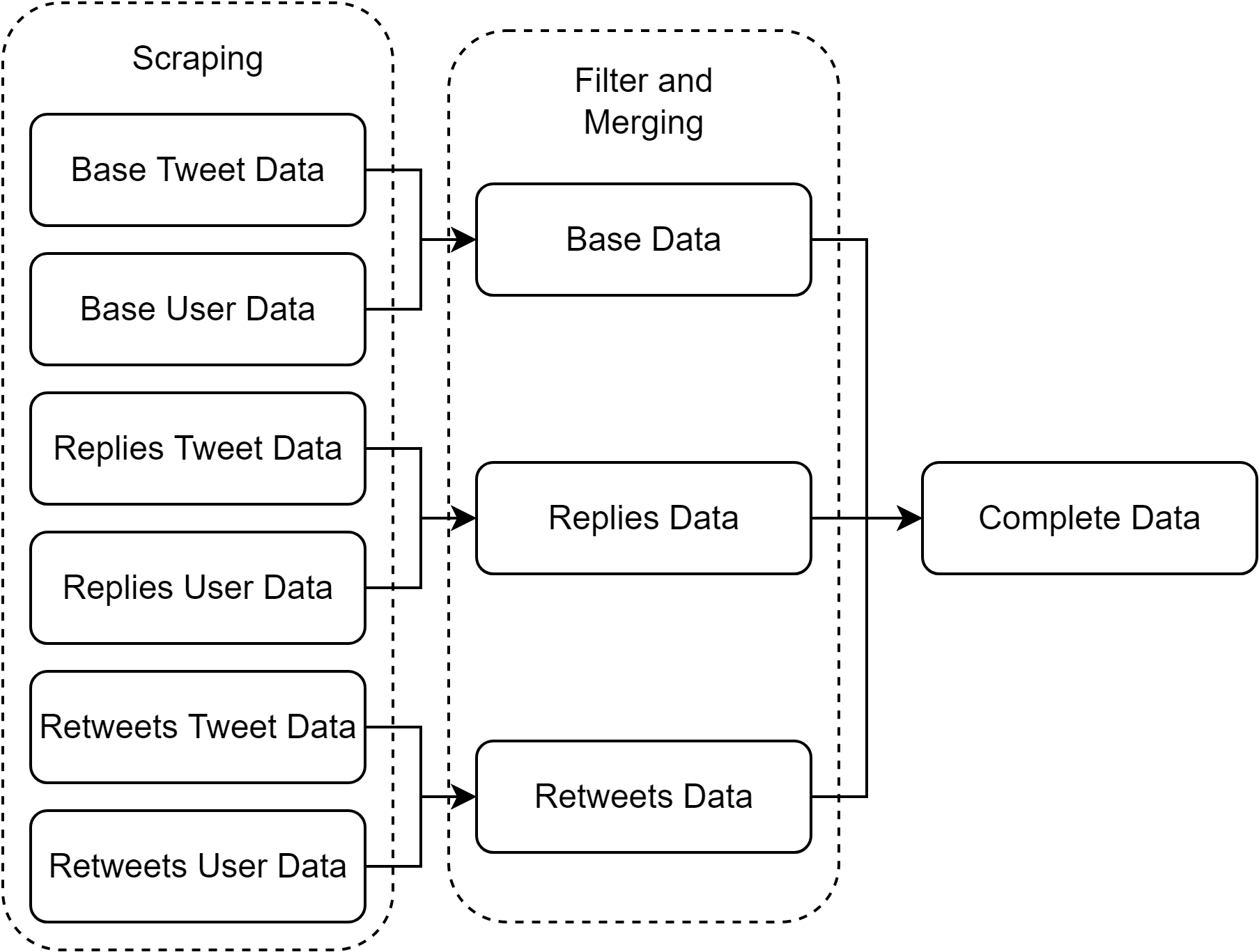
## 2.3 Pengumpulan Data

Pengumpulan data diperoleh berdasarkan *web scraping* pada akun twitter target, dengan menggunakan tools *snscrape*. Dalam penelitian ini, terdapat dua target akun yang dijelaskan pada Tabel 2.1 berikut

**Tabel 2.1** Informasi Akun Target

|  |  |  |
| --- | --- | --- |
| Nama Akun | Posisi | Bergabung sejak |
| \_\_bungs | Akun yang menyerang oposisi | Oktober 2012 |
| catatanali07 | Akun yang menyerang pemerintah | Mei 2019 |

Terdapat beberapa tahapan yang dilalui untuk mendapatkan data mentah yang dapat digunakan, yaitu tahap scraping, filter, dan merging data seperti terdapat pada Gambar 2.2 berikut

[](https://app.diagrams.net/?page-id=C5RBs43oDa-KdzZeNtuy&scale=auto#G1qLIsltYGSbXLyW8bzNjxvnJldMOJTBId)

**Gambar 2.2** Proses *web scraping*

Terdapat tiga jenis tweet yang kami gunakan dalam penelitian, yakni *base* atau *normal tweet*, *replies tweet*, dan *retweets tweet*. Setiap jenis tweet tersebut kami bagi terlebih dahulu ke dalam *tweet data* dan *user data* untuk memudahkan dan mempercepat proses *scraping*. Setiap proses tersebut menghasilkan data mentah yang belum dilakukan filtering fitur-fiturnya. Selain itu juga, *user data* juga perlu untuk digabung dengan *tweet data* untuk dapat mengetahui pengaruh dan informasi lainnya pada data tweet.

Dalam tahapan ini juga, *query* yang digunakan dalam scraping untuk masing-masing jenis tweet dijelaskan pada Tabel 2.2 sebagai berikut

**Tabel 2.2** *Query* yang digunakan pada *scraping*

|  |  |  |
| --- | --- | --- |
|  | \_\_bungs | catatanali07 |
| base tweet | from:\_\_bungs -filter:replies | from:catatanali07 -filter:replies |
| replies tweet | to:\_\_bungs filter:replies | to:catatanali07 filter:replies |
| retweets tweet | to:\_\_bungs filter:nativeretweets | to:catatanali07 filter:nativeretweets |

Dari proses scraping tersebut, diperoleh sejumlah data dengan rincian pada Tabel 2.3 sebagai berikut

**Tabel 2.3** Jumlah data hasil scraping

|  |  |  |
| --- | --- | --- |
|  | \_\_bungs | catatanali07 |
| base tweet | 30580 | 183 |
| replies tweet | 37365 | 1026 |
| retweets tweet | 5 | 0 |

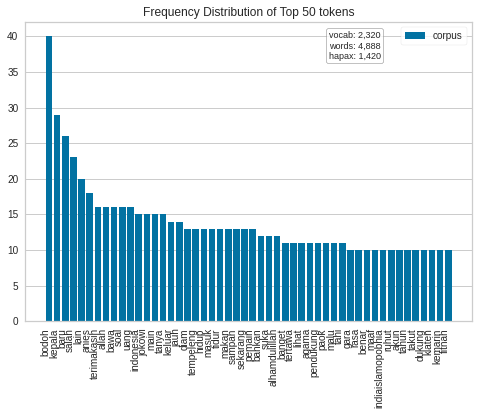
Kemudian dilakukan *filtering* sejumlah 15 kolom, yaitu kolom Id, *Tweet Time*, *Url*, *Tweet Text*, *Likes Count*, *Reply Count*, *Retweet Count*, *Retweeted Text*, *User Id*, *User Name*, *User Follower Count*, *User Following* *Count*, *User Tweets Count*, *User Likes Count*, dan *User Location*. Lalu dilakukan penggabungan data untuk setiap akun dan diperoleh 67950 data untuk akun \_\_bungs dan 1209 data untuk akun catatanali07.

## 2.4 Pre-processing

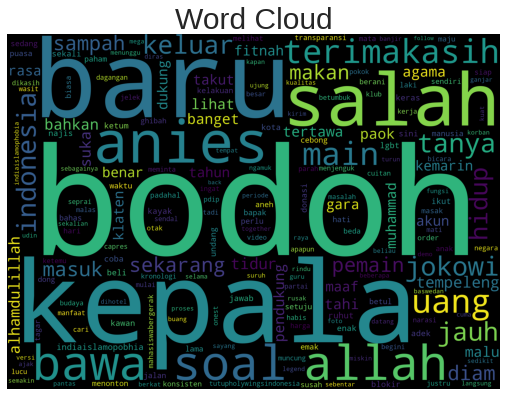
*Pre-processing* merupakan sebuah langkah penting dalam Data Mining untuk membuat data mentah menjadi data yang berkualitas. Data perlu dilakukan *preprocessing* karena dalam data mentah terdapat data yang tidak lengkap, *noise*, dan tidak konsisten. *Pre*-*processing* dilakukan dahulu supaya eksplorasi data akan lebih mudah. Langkah pertama yang dilakukan pada *pre-processing* adalah menghapus *URL*, *mention*, tanda baca, angka, *HTML tag*, dan *hashtag* pada teks *tweet*. Kemudian, teks tersebut di tokenisasi dan huruf kapital diubah menjadi huruf kecil. Langkah terakhir pada *pre-processing* adalah mengubah kata *slang* menjadi kata baku, menghapus *stopword*, dan menghapus kata dengan panjang kurang dari tiga huruf.

## 2.5 Eksplorasi Data

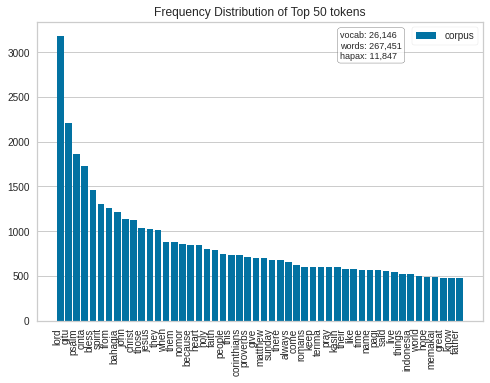
Eksplorasi Data dilakukan dengan membuat dua visualisasi, yaitu visualisasi distribusi frekuensi token dan *word cloud*.

**

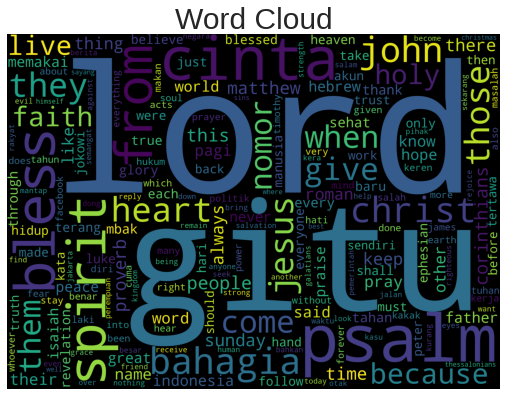
**Gambar 2.1** Visualisasi Distribusi Frekuensi Token *Tweet* Akun “@Catatanali\_07”

****

**Gambar 2.2** *Word Cloud Tweet* Akun “@Catatanali\_07”

****

**Gambar 2.3** Visualisasi Distribusi Frekuensi Token *Tweet* Akun “@\_\_Bungs”



**Gambar 2.4** *Word Cloud Tweet* Akun “@\_\_Bungs”

## 

## 2.6 Ekstraksi Fitur

Untuk mendapatkan hasil *clustering* yang terbaik, ekstraksi fitur dilakukan dengan tiga metode untuk kemudian dicari metode yang terbaik. Metode yang digunakan adalah Word2vec, TF-IDF, dan CountVectorizer. Pada metode Word2vec, model Word2vec dibuat dari hasil *pre-processing* teks *tweet* kemudian pada setiap *tweet*, setiap kata dihitung *word vector* dan dihitung rata-rata dari *word vector* pada tweet tersebut untuk mendapatkan *sentence vector*. Pada metode TF-IDF dan CountVectorizer, data teks *tweet* setelah *pre-processing* ditransformasikan menggunakan TfidfVectorizer dan CountVectorizer.

## 2.7 Metode Clustering

Hasil dari tahap ekstraksi fitur digunakan untuk melakukan *clustering*. Metode *clustering* yang digunakan adalah *K-Means Clustering*. Parameter jumlah *cluster* ditentukan berdasarkan *silhouette score* yang terbaik atau menggunakan *elbow method*.

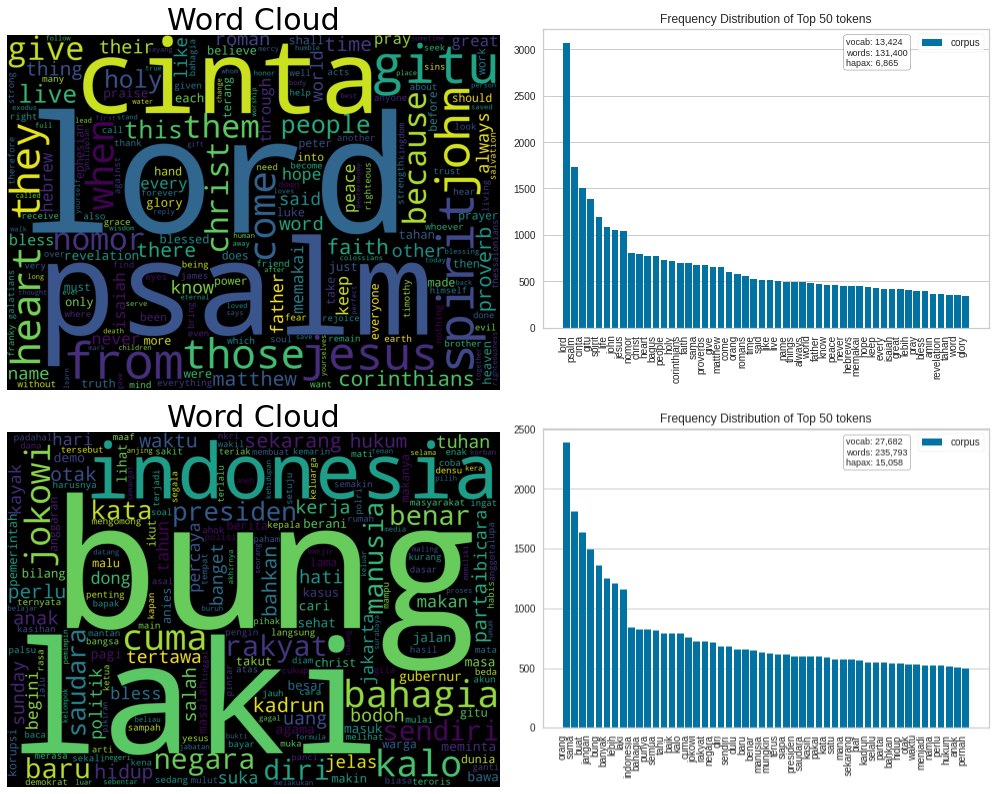
# 

# III. HASIL DAN PEMBAHASAN

## 3.1 Word2vec

Setelah melakukan *clustering* menggunakan metode Word2vec, *silhouette score* tertinggi adalah 0.4618 dengan jumlah *cluster* 2.

|  |  |
| --- | --- |
| Jumlah *cluster* | *Silhouette score* |
| 2 | 0.4618 |
| 3 | 0.2507 |
| 4 | 0.1726 |
| 5 | 0.1434 |
| 6 | 0.1246 |
| 7 | 0.1088 |
| 8 | 0.1068 |
| 9 | 0.1007 |
| 10 | 0.0912 |



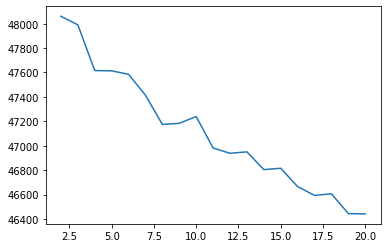
**Gambar 3.1** *Word Cloud* dan Distribusi Frekuensi *Clustering Tweet* Akun “@\_\_Bungs” Menggunakan Word2vec

Perbedaan yang paling jelas antara *cluster* pertama dengan yang kedua terdapat pada kata berbahasa Inggris dimana *cluster* pertama memiliki kata berbahasa *Inggris* yang jauh lebih banyak daripada *cluster* kedua. Metode ini tidak dapat digunakan untuk melakukan klasifikasi *dangerous speech*.

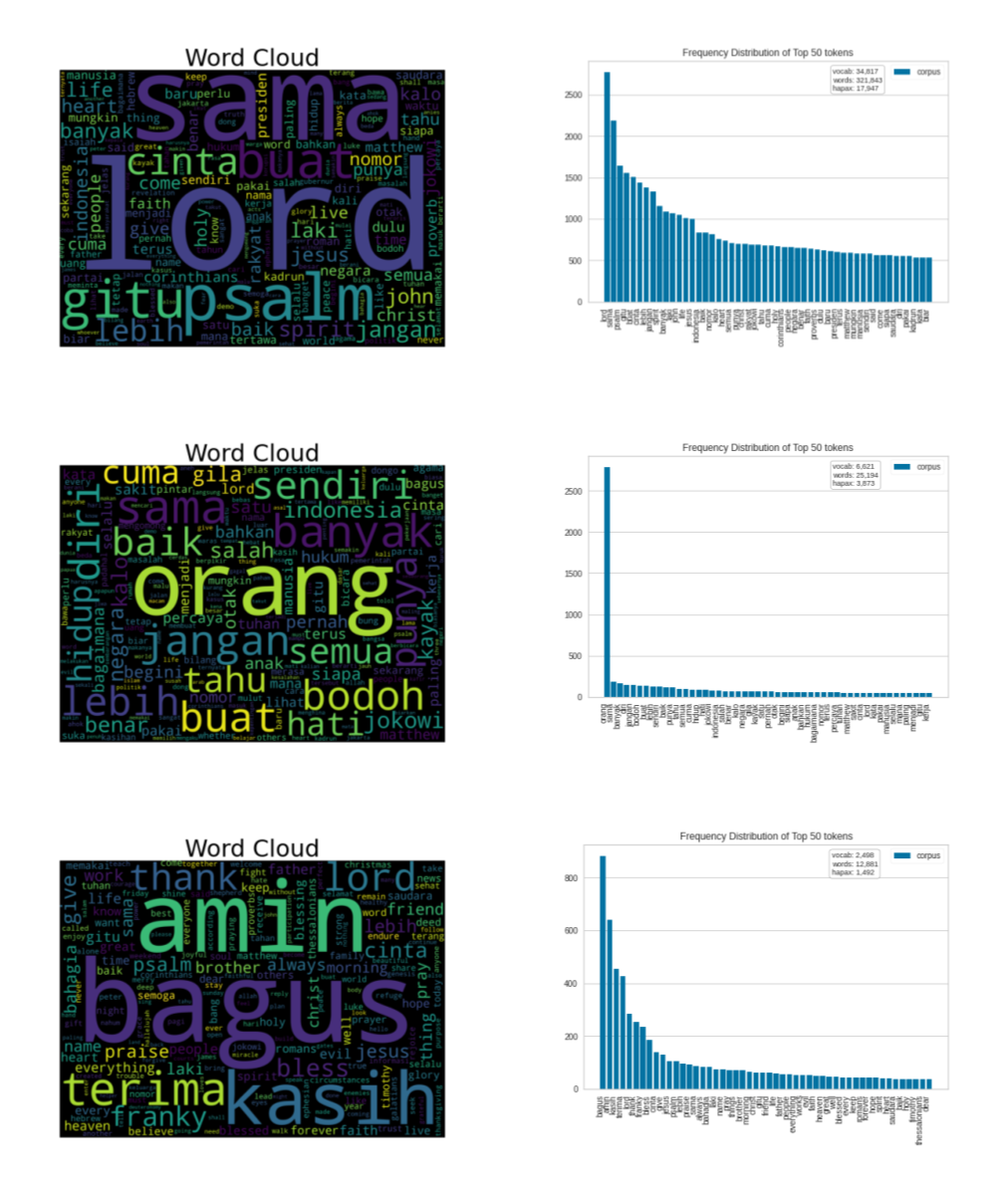
## 3.2 TF-IDF

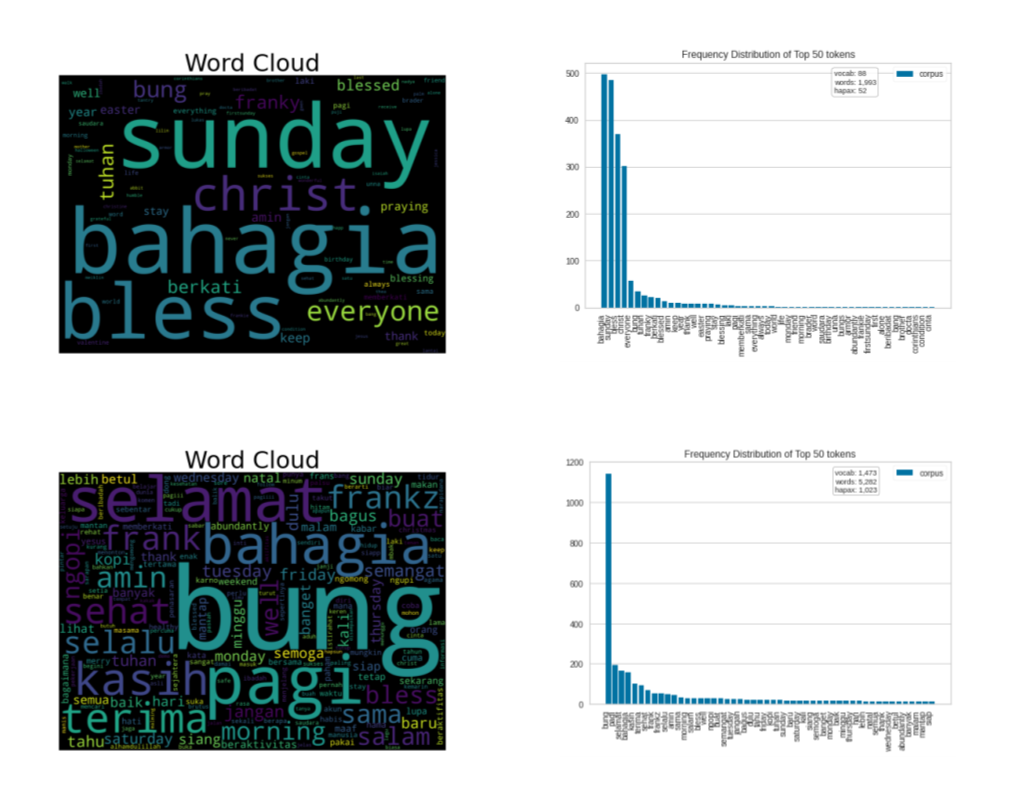
Setelah melakukan *clustering* menggunakan metode TF-IDF, semakin banyak jumlah *cluster*, semakin naik *silhouette score*. Grafik *inertia* juga tidak menunjukkan adanya *elbow*, sehingga untuk memudahkan analisis *clustering*, jumlah *cluster* yang digunakan adalah 5.

|  |  |
| --- | --- |
| Jumlah *cluster* | *Silhouette score* |
| 2 | 0.0026 |
| 3 | 0.0026 |
| 4 | 0.0029 |
| 5 | 0.0046 |
| 6 | 0.0056 |
| 7 | 0.0063 |
| 8 | 0.0075 |
| 9 | 0.0076 |
| 10 | 0.0081 |



**Gambar 3.2** Grafik *Inertia Cluster* *Tweet* Akun “@\_\_Bungs” Menggunakan TF-IDF





**Gambar 3.3** *Word Cloud* dan Distribusi Frekuensi *Clustering* *Tweet* Akun “@\_\_Bungs” Menggunakan TF-IDF

*Cluster* pertama mengandung jumlah *vocab* dan kata yang terbanyak. *Cluster* tersebut mengandung kata berupa buku yang pada Alkitab seperti “psalm”, “john”, “corinthians”, “proverbs”, dan “matthew”. *Clustering* menggunakan TF-IDF memiliki *silhouette score* yang sangat rendah dan setiap *cluster* tidak memiliki ciri-ciri yang membedakan satu sama yang lain. *Cluster* yang mengandung *tweet* yang dapat mengandung *dangerous speech* adalah *cluster* kedua karena terdapat kata-kata seperti “bodoh”, “jokowi”, “otak”, “hukum”, dan “gila”.

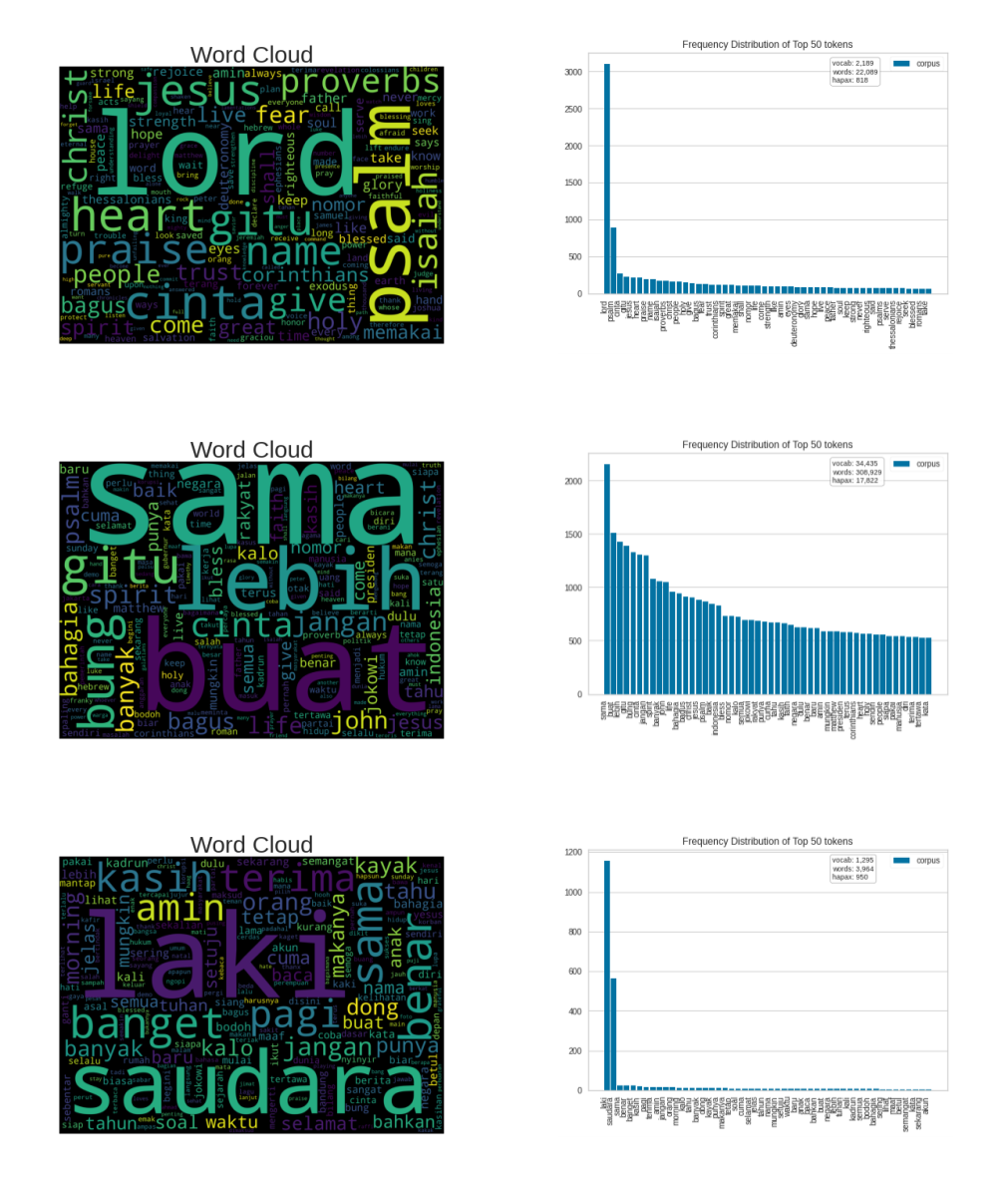
## 3.3 CountVectorizer

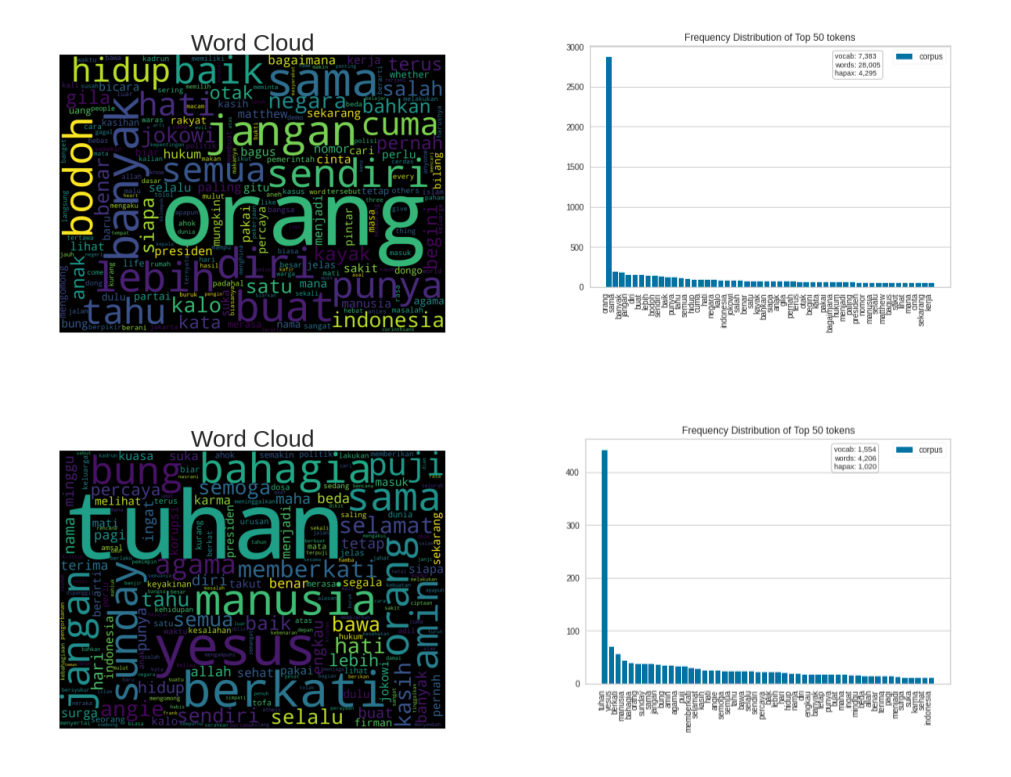
Setelah melakukan *clustering* menggunakan metode CountVectorizer, semakin banyak jumlah *cluster*, semakin naik *silhouette score*. Grafik *inertia* juga tidak menunjukkan adanya *elbow*, sehingga untuk memudahkan analisis *clustering*, jumlah *cluster* yang digunakan adalah 5.

|  |  |
| --- | --- |
| Jumlah *cluster* | *Silhouette score* |
| 2 | 0.0103 |
| 3 | 0.0142 |
| 4 | 0.0161 |
| 5 | 0.017 |
| 6 | 0.0175 |
| 7 | 0.0175 |
| 8 | 0.0179 |
| 9 | 0.0168 |
| 10 | 0.0198 |

# 

**Gambar 3.4 Grafik *Inertia Cluster* *Tweet* Akun “@\_\_Bungs” Menggunakan CountVectorizer**

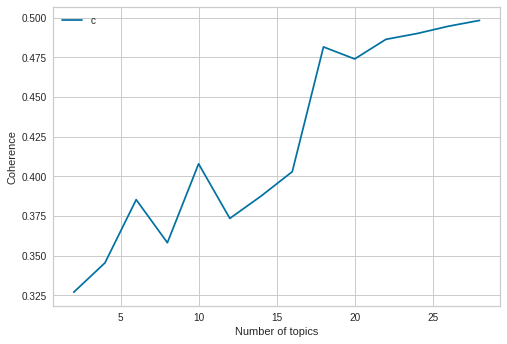




**Gambar 3.5** *Word Cloud* dan Distribusi Frekuensi *Clustering* *Tweet* Akun “@\_\_Bungs” Menggunakan CountVectorizer

*Cluster* kelima paling berhubungan dengan agama dengan adanya kata seperti “tuhan”, “yesus”, “berkati”, dan “agama”. *Cluster* pertama juga membahas agama, tetapi hal yang dibahas mengarah ke Alkitab dengan kata seperti “psalm”, “isaiah”, dan “proverbs”. *Cluster* keempat membahas politik Indonesia dengan kata “negara”, “indonesia”, “jokowi”, dan “presiden”. *Cluster* keempat juga memiliki kata yang bersifat negatif seperti “bodoh” dan “gila”, sehingga *cluster* ini mengandung *tweet* yang dapat digolongkan sebagai *dangerous speech*.

## 3.4 LDA *Topic Modelling*

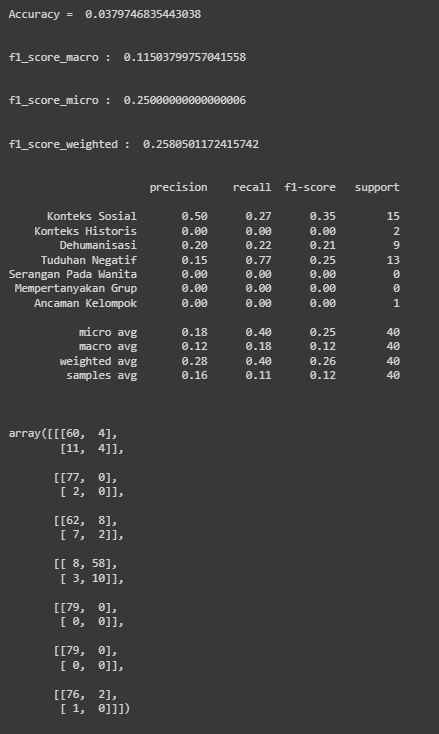


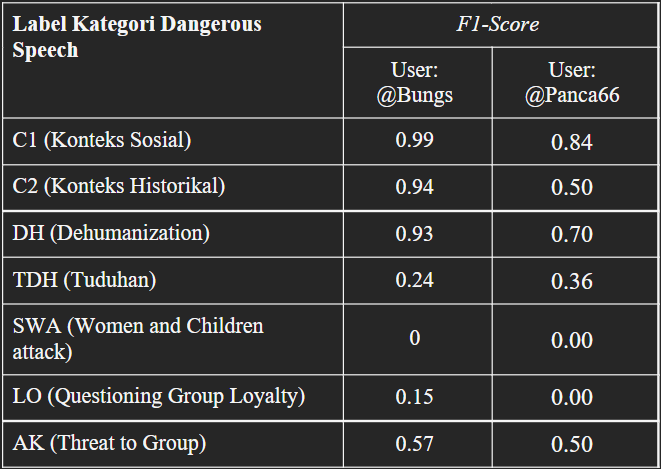


Berdasarkan grafik *coherence value*, jumlah topik yang terbaik adalah 18 karena *coherence value* dengan jumlah topik 18 terdapat pada puncak grafik. Dari hasil pembagian topic tersebut, berikut analisa dari masing-masing topik

|  |  |  |
| --- | --- | --- |
| Topik | Contoh kata | Analisis |
| 4 | partai, kata, kerja, suka, semakin, cara, mati, arti, tinggi, jauh | Kluster ini rata-rata berisi dangerous speech  Contoh :  “@SantorinisSun Gw semakin yakin. Syarat pertama untuk jadi kadrun adalah: HARUS BEGO” |
| 10 | gubernur, sebentar, pikiran, densus, urusan, mental, negatif, maju, tanah, senang | Pilgub Jakarta |

**3.5 Klasifikasi *dangerous speech* akun “@catatanali07”**

****

****

Klasifikasi dilakukan dengan training data dari Bu Yulian dan test data dari manual *labeling* dari *tweet* akun “catatanali07”. *Preprocessing* pada data teks sama seperti pada *clustering* dan hasil *preprocessing* ditransformasikan menjadi TFIDF. Metode klasifikasi yang digunakan adalah BaggingClassifier. Akurasi yang diperoleh adalah 3.80%, nilai makro F1 adalah 11.50%, nilai mikro F1 adalah 25%, dan nilai *weighted* F1 adalah 25.81%. Model klasifikasi kurang akurat terutama pada kelas Dehumanisasi dan Tuduhan Negatif.

# 

# 

# IV. KESIMPULAN

Dengan mengubah teks menjadi *token*, data teks *tweet* dapat diubah menjadi vektor yang bisa digunakan untuk melatih suatu model untuk mengelompokkan *tweet* menjadi beberapa *cluster*. *Silhouette score* tertinggi didapatkan saat mendapatkan vektor teks menggunakan Word2vec dan menggunakan jumlah *cluster* 2. Hasil *cluster* menggunakan Word2vec menghasilkan *cluster* pertama yang mengandung banyak *tweet* yang berbahasa Inggris dan *cluster* kedua yang mengandung banyak *tweet* yang berbahasa Indonesia. Oleh karena itu, *feature extraction* Word2vec dengan metode *K-Means Clustering* tidak dapat digunakan untuk melakukan klasifikasi *dangerous speech*. Metode *feature extraction* yang dapat digunakan untuk mengklasifikasi *tweet* yang dapat mengandung *dangerous speech* adalah *CountVectorizer*. *Silhouette score* pada metode ini untuk jumlah *cluster* 5 adalah 0.017 dan terdapat satu *cluster* yang mengandung *tweet* yang dapat mengandung *dangerous speech*.

# DAFTAR PUSTAKA

Sirisuriya, SCM de S. 2015. Proceedings of 8th International Research Conference. A Comparative Study on Web Scraping.

Assuja, Maulana Aziz. 2016. Jurnal Teknoinfo. Analisis Sentimen Tweet Menggunakan Backpropagation Neural Network.

Devi Putri Isnarwaty, Irhamah. 2019. Jurnal Sains dan Seni. Text Clustering pada Akun TWITTER Layanan Ekspedisi JNE, J&T, dan Pos Indonesia Menggunakan Metode Density-Based Spatial Clustering of Applications with Noise (DBSCAN) dan K-Means.

Beryandhi, M. (2020, Oktober 27). Media Baru dan Fenomena Hate Speech di Indonesia: Media Penyakit Sosial Baru. Diakses 16 Januari 2022, dari <https://kumparan.com/mohamad-beryandhi/media-baru-dan-fenomena-hate-speech-di-indonesia-media-penyakit-sosial-baru-1uTKOats1y9>

Muthahharah, Isma dan Agusalim Juhari. 2021. Jurnal Varian. Cluster Analysis with Complete Linkage and Ward's Method for Health Service Data in Makassar City.