**PROPOSAL PRA SKRIPSI**

**ANALISA TOPIK DATA MEDIA SOSIAL TWITTER TERHADAP**

**PEJABAT PUBLIK MENGGUNAKAN LATENT SEMANTIC ANALYSIS**

**(Studi Kasus : Gubernur Jawa Tengah)**



**Disusun Oleh :**

**WIDYA SULISTYANI**

**Nomor Mahasiswa : 165410086**

**PROGRAM STUDI TEKNIK INFORMATIKA**

**SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER AKAKOM YOGYAKARTA**

**2020**

# 

# HALAMAN PERSETUJUAN

|  |  |  |
| --- | --- | --- |
| **Judul** | **:** | **Analisa Topik Data Media Sosial Twitter Terhadap Pejabat Publik Menggunakan Latent Semantic Analysis** |
| **Nama** | **:** | **Widya Sulistyani** |
| **Nomor Mhs** | **:** | **165410086** |
| **Program Studi** | **:** | **Teknik Infomatika** |
| **Jenjang** | **:** | **Strata Satu (S1)** |
| **Tahun** | **:** | **2020** |

**Telah diperiksa dan disetujui**

**Yogyakarta, 17 Juni 2020**

**Mengetahui**

**Dosen Pembimbing**

**[Dini Fakta Sari, S.T., M.T.](https://scholar.google.co.id/citations?user=RQCxqBQAAAAJ&hl=id)**

# DAFTAR ISI

Table of Contents

[HALAMAN PERSETUJUAN i](#_Toc43791735)

[DAFTAR ISI ii](#_Toc43791736)

[DAFTAR GAMBAR iii](#_Toc43791737)

[DAFTAR TABEL iv](#_Toc43791738)

[BAB 1 PENDAHULUAN 1](#_Toc43791739)

[1.1 Latar Belakang 1](#_Toc43791740)

[1.2 Rumusan Masalah 2](#_Toc43791741)

[1.3 Ruang Lingkup 2](#_Toc43791742)

[1.4 Tujuan Penelitian 3](#_Toc43791743)

[1.5 Manfaat Penelitian 3](#_Toc43791744)

[BAB 2 TINJAUAN PUSTAKA DAN DASAR TEORI 4](#_Toc43791745)

[2.1 Tinjauan Pustaka 4](#_Toc43791746)

[2.2 Dasar Teori 8](#_Toc43791747)

[BAB 3 METODE PENELITIAN 14](#_Toc43791748)

[3.1 Data 14](#_Toc43791749)

[3.2 Peralatan 14](#_Toc43791750)

[3.3 Prosedur dan Pengumpulan Data 15](#_Toc43791751)

[3.4 Analisis dan Perancangan Sistem 15](#_Toc43791752)

[3.5 Analisis dan Rancangan Sistem 17](#_Toc43791753)

[3.6 Prosedur dan Pengumpulan Data 17](#_Toc43791754)

[Text Preprocessing BY INFORMATIKALOGI · PUBLISHED NOVEMBER 27, 2016 · UPDATED JULY 11, 2017 https://informatikalogi.com/text-preprocessing/ 18](#_Toc43791755)

# DAFTAR GAMBAR

# DAFTAR TABEL

[Tabel 2.1 Perbandingan Penelitian 7](#_Toc43791711)

# PENDAHULUAN

## Latar Belakang

Media sosial merupakan sebuah *platform* yang penggunanya dapat berbagi pandangan, ide, perasaan, dan emosi mengenai suatu konteks peristiwa atau hal yang sedang terjadi saat itu. Media sosial semakin populer seiring dengan semakin mudahnya akses internet pada laptop, komputer, dan perangkat *mobile*. Banyaknya pengguna media sosial yang aktif menyuarakan pendapat mereka, menjadikan media sosial sebagai sumber data utama dalam bidang *social data mining* (Matwin 2013).

Twitter merupakan salah satu media sosial populer di dunia sejak pertama kali dipublikasikan pada tahun 2006 (Singh 2015). Indonesia menggunakan Twitter sejak awal dipublikasikan dan termasuk dalam pengguna Twitter yang paling produktif. Pada tahun 2014 terdapat sekitar 29 juta pengguna Twitter di Indonesia, sehingga Indonesia menempati peringkat kelima sebagai negara dengan *tweet* terbanyak (Carley *et al* 2015).

Pengguna twitter dapat memanfaatkan twitter untuk berbagi informasi untuk berbagai kebutuhan seperti kebutuhan menyampaikan pendapat, kebutuhan bisnis hingga kebutuhan publik. Tweet yang dibagikan oleh pengguna twitter memiliki beragam topik. Topik mewakili isi dari banyak tweet yang memiliki konteks yang sama. Dari tweets tersebut dapat ditemukan topik utama yang banyak diperbincangkan dengan melakukan analisis topik. Analisis topik dilakukan untuk memberikan gambaran umum yang mewakili topik utama. Banyaknya tweet yang ada, diperlukan metode yang lebih cepat dan efektif yaitu dengan metode pemodelan topik.

Pemodelan topik dapat digambarkan sebagai metode untuk menemukan kelompok kata (topik) dari kumpulan dokumen yang dapat merepresentasikan dengan baik informasi yang ada dalam kumpulan dokumen tersebut (Nair 2016). Salah satu model dari pemodelan topik adalah model *Latent Semantic Analysis*.

## Rumusan Masalah

Berdasarkan latar belakang diatas, maka rumusan masalah pada penelitian ini adalah :

### Bagaimana membuat sistem yang mampu melakukan pengambilan data twitter.

### Bagaimana memproses data twitter untuk menemukan informasi berupa topik yang tersembunyi dan dapat menampilkan trending topik menggunakan pemodelan topik *Latent Semantic Analysis*.

## Ruang Lingkup

Ruang lingkup dari penelitian ini adalah sebagai berikut :

### Data twitter yang digunakan diperoleh dari proses *crawling* menggunakan keyword “ganjar pranowo” selaku gubernur jawa tengah.

### Penelitian dilakukan pada rentang waktu 1 maret 2020 sampai 1 juni 2020.

### Korpus yang digunakan adalah korpus berbahasa Indonesia.

### Pemodelan topik yang digunakan adalah *Latent Semantic Analysis* (LSA).

## Tujuan Penelitian

Penelitian ini diharapkan dapat memberikan pengetahuan tentang bagaimana cara melakukan eksplorasi opini dan menganalisa data twitter sehingga dapat memberikan pengetahuan tentang topik yang tersembunyi pada jejaring sosial twitter terhadap pejabat publik.

## Manfaat Penelitian

Manfaat dari penelitian ini adalah :

### Bagi Ganjar Pranowo selaku pejabat publik, penelitian ini dapat membantu mengefisienkan waktu untuk mengetahui isu yang berkembang dan dapat mengetahui aspirasi masyarakat, sehingga dapat menjadi bahan informasi dan bahan evaluasi bagi Ganjar Pranowo selaku pejabat publik.

### Bagi peneliti lain, penelitian ini dapat menjadi bahan referensi untuk penelitian text mining, utamanya pemodelan topik menggunakan *latent semantic analysis*.

# TINJAUAN PUSTAKA DAN DASAR TEORI

## Tinjauan Pustaka

Untuk membedakan penelitian ini dengan penelitian sebelumnya maka penulis melakukan studi pustaka terhadap peneliti peneliti terdahulu.

Ardytha Luthfiarta, Junta Zeniarja & Abu Salam (2013). Peneliti melakukan peringkasan dokumen otomatis dengan menggunakan algoritma *Latent Semantic Analysis* (LSA). Berdasarkan percobaan-percobaan yang telah dilakukan dapat disimpulkan bahwa Peringkas Dokumen Otomatis dengan *Latent Semantic Analysis* (LSA) pada Proses Clustering Dokumen Teks Berbahasa Indonesia dapat meningkatkan kinerja clustering dokumen lebih baik dari pada Peringkas Dokumen Otomatis dengan Metode Fitur dan Proses Clustering Dokumen Standar, mengalami peningkatan dari tingkat akurasi 65,92 % untuk proses clustering standar menjadi 71,04% untuk proses clustering dokumen menggunakan peringkas dokumen otomatis dengan *Latent Semantic Analysis* (LSA).

Dimas Bagus C. W. (2017). Peneliti membuat aplikasi dengan metode metode text mining untuk melakukan analisis sentimen untuk menemukan topik topik yang terdapat pada data tweets dan melihat keterkaitan antar kata pada khasus pilkada DKI Putaran 2.

Keke Putri Utami (2017) melakukan pemodelan topik menggunakan *Latent Dirichlet Allocation* pada lima lokasi *tweet* di Kota Bogor dan rentang waktu tertentu, dan berhasil membentuk topik dengan informasi atau deskripsi topik untuk setiap lokasi *tweet*. Deskripsi topik untuk setiap lokasi *tweet* menunjukkan topik-topik tersebut sedang banyak dibicarakan oleh pengguna Twitter pada rentang waktu yang ditentukan. Jumlah topik yang ditentukan untuk data csv setiap lokasi *tweet* telah dapat menghasilkan kumpulan kata yang membentuk topik dengan baik. Informasi yang mewakili isi topik dapat dimanfaatkan oleh pembaca atau *stakeholder* terkait dalam memahami setiap perkembangan isu terkini.

Nurina Savanti Widya Gotami, Indriati, Ratih Kartika Dewi (2018).Peneliti melakukan peringkasan teks otomatis secara ekstraktif pada artikel berita kesehatan berbahasa Indonesia dengan menggunakan LSA dapat diterapkan dengan cara LSA sebagai algoritme untuk mendapat kalimat kalimat yang memiliki keterkaitan kata dengan pendekatan secara semantik dengan menggunakan SVD sebagai fitur penghilang redudansi atau *noise* pada kata tertentu. Serta penggunaan *Cross method* LSA sebagai pengekstrasi ringkasan yang akan dipilih dari artikel berita kesehatan dalam data teks dokumen.

Tinjauan keenam ditulis oleh Septian Narsa Putra (2018). Peneliti membuat sistem yang dapat melakukan klasifikasi terhadap berita yang diupload pada akun twitter Divis Humas Polri ke dalam tiga katagori yaitu berita kegiatan polisi, komentar masyarakat dan layanan masyarakat selama empat tahun ke belakang. Setelah berita diklasifikasi kemudian dicari sentimen dari setiap topik. Metode yang digunakan adalah Naive Bayes Classifier.

Pada penelitian kali ini akan dibuat sistem yang dapat melakukan clustering berdasarkan kesamaan topik pada akun twitter pejabat publik, sehingga dapat ditemukan topik yang tersembunyi dari tweets yang ada menggunakan *latent semantic analysis*.

Perbandingan dari penelitian - penelitian diatas dapat dilihat pada table 2.1

**Tabel 2. 1 Perbandingan Penelitian**

|  |  |  |  |
| --- | --- | --- | --- |
| Penulis | Objek | Metode | Hasil |
| Ardytha Luthfiarta, Junta Zeniarja & Abu Salam (2013) | Dokumen Teks Berbahasa Indonesia | *Latent Semantic Analysis (LSA)* | Sistem untuk proses c*lustering* dokumen teks berbahasa Indonesia. |
| Dimas Bagus C. W. (2017) | Data Twitter | Supervised Learning : Naïve Bayes dan Unsupervised Learning : Association Rule, K Means & Topic Modeling | Aplikasi dengan menggunakan metode metode text mining untuk melakukan analisis sentiment dan menemukan topik topik yang terdapat pada data tweets. |
| Keke Putri Utami (2017) | Data Twitter | *Latent Dirichlet Allocation* | Sistem pemodelan topik LDA pada 5 lokasi tweet di kota bogor membentuk informasi topik setiap lokasi tweet. |
| Nurina Savanti Widya Gotami 1, Indriati 2, Ratih Kartika Dewi 3 (2018) | Artikel berita kesehatan dalam data teks dokumen | *Latent Semantic Analysis (LSA)* | Sistem peringkasan teks otomatis secara ekstraktif pada artikel kesehatan berbahasa Indonesia. |
| Septian Narsa Putra (2018) | Data Twitter | Naive Bayes Classifier | Sistem klasifikasi berita pada akun twitter. |
| Widya Sulistyani (2020) | Data Twitter | *Latent Semantic Analysis (LSA)* | Sistem yang dapat menemukan topik tersembunyi di dalam tweets menggunakan pemodelan topik Latent Semantic Analysis. |

## Dasar Teori

### Twitter

Twitter merupakan salah satu media sosial dengan layanan *microbloging* yang terkenal dan memungkinkan para penggunanya untuk menulis sesuatu atau yang biasa disebut *tweet*. Twitter digunakan untuk mengutarakan opini publik maupun berita resmi dari suatu instansi atau dari pejabat publik. Twitter dibangun oleh Jack Dorsey pada tahun 2006 dengan alamat http://www.twitter.com, jika seseorang ingin menggunakan twitter seseorang harus terlebih dahulu memiliki akun, untuk registrasinya dapat dilakukan pada alamat tersebut. Pengguna dapat menulis pesan berdasarkan topik dengan tanda #(tagar). Sedangkan untuk menyebut atau membalas pesan dari pengguna lain bisa menggunakan tanda @(diikuti nama akun yang akan dibalas).

### Text Mining

Text mining memiliki definisi menambang data yang berupa teks dimana sumber data biasanya di dapatkan dari dokumen, dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen.

Dalam proses penambangan data terdapat proses pengubahan data menjadi data yang terstruktur sesuai dengan kebutuhan, proses ini disebut *Text Prepocessing*, dengan tahapannya adalah *Case Folding, Tokenizing* dan *Filtering*. *Case Folding* adalah proses mengubah semua huruf dalam tweet menjadi huruf kecil (lowercase). Hanya huruf “a” sampai dengan “z” yang diterima dan karakter selain huruf dihilangkan. Tahapan *Tokenizing* adalah proses pemotongan kalimat (*string*) berdasarkan kata peyusunnya. Kemudian tahap *filtering*, merupakan tahapan mengambil kata - kata penting dari hasil token dan membuang sebagian kata tertentu.

### Perhitungan TF-IDF

### Tahap perhitungan TF-IDF dilakukan guna memperoleh bobot dari setiap kata di setiap tweet. Pembobotan pada setiap kata dilakukan menggunakan term indeks dari *preprocessing* dan akan dihasilkan bobot dari sebuah kata yang didasari oleh jumlah frekuensi kemunculan kata pada tweet. Pembobotan kata pada tiap kata diperoleh dari perhitungan frekuensi kemunculan term dalam tweet yang disebut TF (Term Frequency), lalu menghitung frekuensi kemunculan tweet yang mengandung term yang disebut DF (document frequency) serta perhitungan IDF (inverse Document Frequency) yang menghitung jumlah tweet yang mengandung term yang dicari dari tweet yang ada.

CodeCogsEqn (2) (1)

CodeCogsEqnCodeCogsEqn (1) (2)

(3)

Keterangan :

TF-IDF = bobot tweet terhadap kata

TF = banyak kata dicari pada sebuah tweet

IDF = Inversed Document Frequency

N = total tweet

### Pemodelan Topik

*Topic Modeling* atau pemodelan topik merupakan metode *clustering* yang termasuk dalam *unsupervised learning*. Dalam *unsupervised learning* tidak ada label untuk suatu objek. Terdapat 3 tipe *clustering* yaitu *hard clustering*, *hierarchical clustering*, dan *soft/fuzzy clustering*. Pemodelan topik termasuk dalam *soft/fuzzy clustering* yang mana setiap objek dapat dimiliki lebih dari satu *cluster* dengan tingkat tertentu. Sebagai contoh, perhitungan kemungkinan yang dimiliki objek untuk tergabung dalam suatu *cluster* adalah berbeda (Doig 2015).

Setiap harinya sejumlah besar data terkumpul dan menjadikan semakin banyaknya informasi yang tersedia. Pemodelan topik menyediakan suatu metode untuk mengatur, memahami, dan meringkas kumpulan dokumen. Pemodelan topik membantu dalam hal menemukan pola topik tersembunyi yang ada dalam kumpulan dokumen, memberikan keterangan dokumen sesuai dengan topik, memanfaatkan pemberian keterangan ini untuk mengatur, mencari, dan meringkas data teks. Pemodelan topik dapat digambarkan sebagai metode untuk menemukan kelompok kata (topik) dari kumpulan dokumen yang dapat merepresentasikan dengan baik informasi yang ada dalam kumpulan dokumen tersebut(Nair 2016). Terdapat beberapa teknik yang dapat digunakan untuk pemodelan topik, salah satunya adalah pemodelan topik *Latent Semantic Analysis*.

### *Latent Semantic Analysis*

### LSA *(Latent Semantic Analysis)* adalah metode statistik aljabar yang mengekstrak struktur semantik yang tersembunyi dari kata dan kalimat, untuk mencari interelasi diantara kalimat dan kata, digunakan metode aljabar Singular Value Decomposition (SVD). Disamping mempunyai kapasitas relasi model diantara kata dan kalimat, SVD ini mempunyai kapasitas reduksi noise yang membantu untuk meningkatkan akurasi.

Langkah –langkah pada algoritma LSA adalah sebagai berikut :

1. Pembuatan Inputan Matriks

Inputan matriks diperoleh dari proses perhitungan TF-IDF yang membentuk sebuah matriks A.

2. *Singular Value Decomposition*

Metode *singular value decomposition* (SVD) mengubah matriks yang lebih sederhana dengan mendekomposisi menjadi 3 komponen matriks. SVD memiliki kemampuan pemodelan hubungan antara kata - kata dan kalimat serta dapat mereduksi noise yang membantu untuk meningkatkan akurasi.

CodeCogsEqn (4)

Keterangan :

A = Input Matriks

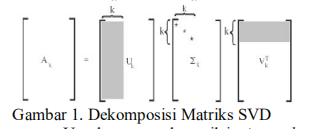
U = Matriks berdimensi m x k

∑ = Matriks diagonal berdimensi k x k

V = Matriks berdomensi k x n

m = Jumlah baris matriks

n = Jumlah kolom matriks



**Gambar 3.1 Dekomposisi Matriks SVD**

Untuk menentukan nilai *eigenvalue* serta *eigenvector* dari matriks N adalah sebagai berikut :

CodeCogsEqn (1) (5)

Keterangan :

N = Matriks bujursangkar dimensi *m* atau *n*

A = Matriks input proses *preprocessing* TF-IDF

AT= Matriks transpose dari inputan TF-IDF

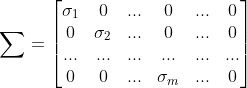
Untuk mencari nilai *eigenvalue* diperoleh dari determinan matriks dengan aturan *Sarrus* sebagai berikut :

CodeCogsEqn (2) (6)

untuk mencari nilai matriks ∑ sebagai berikut :

CodeCogsEqn (3) (7)

Matriks ∑ merupakan matriks diagonal yang terbentuk dari yang diperoleh dari akar *eigenvalue* (λ) hasil pemfaktoran determinan dan menyusun nilai σi sebagai berikut :



Matriks U dibentuk dari perkalian matriks A yang merupakan matriks inputan awal dari proses perhitungan TF-IDF serta matriks V yang diperoleh dari proses normalisasi eigenvector. Sedangkan matriks S-1 merupakan matriks S yang kemudian di invers.

CodeCogsEqn (5) (8)

# METODE PENELITIAN

## Analisis Sistem

### Kebutuhan Data

### Data yang dibutuhkan dalam penelitian ini adalah data *tweets* dengan kata kunci “Ganjar Pranowo” selaku pejabat publik.

### Kebutuhan Proses

Terdapat tiga proses utama dalam pengembangan sistem yang akan dibuat, diantaranya :

1. Input

Pada proses ini diperlukan *tweets* yang dengan kata kunci “ganjar pranowo” selaku pejabat publik.

1. Pengolahan
   1. Sistem melakukan pengambilan data yang berkaitan dengan kata

kunci.

* 1. Dilakukan *preprocessing* untuk membersihkan data tweets, yang

terdiri dari beberapa proses, diantaranya :

*- Case Folding* adalah proses mengubah semua huruf dalam tweet menjadi huruf kecil (lowercase). Hanya huruf “a” sampai dengan “z” yang diterima dan karakter selain huruf dihilangkan.

- *Tokenizing* adalah proses pemotongan kalimat (*string*) berdasarkan kata peyusunnya.

- *Filtering* merupakan tahapan mengambil kata - kata penting dari hasil token dan membuang sebagian kata tertentu yang disebut sebagai *stopword* seperti , “di”, “ke”, “dari” dan “dan”.

* 1. Analisa topik data tweets dan trending topik menggunakan library python.

1. Output

Hasil analisa data tweets berupa topik tersembunyi dan trending topik.

## Peralatan

### Perangkat Lunak

Sistem Operasi Windows

File csv sebagai *database*

Visual Studio Code

Python 3.8.2

### Perangkat Keras

*Processor* Core i3 7020U

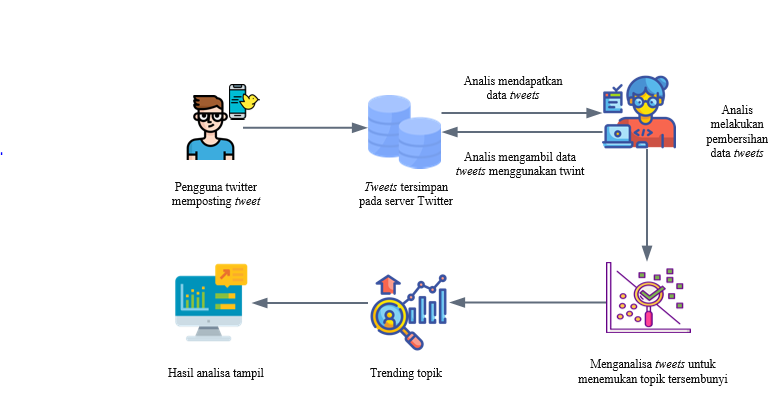
RAM 4 GB

Hardisk 500 GB

## Prosedur dan Pengumpulan Data

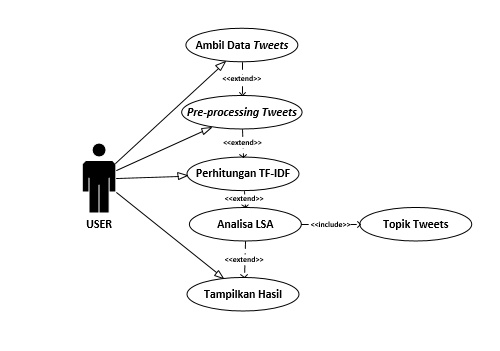
Data yang digunakan pada penelitian ini adalah data tweets yang diambil pada rentang waktu 1 maret sampai 1 juni 2020. Prosedur pengumpulan data pada penelitian ini adalah dengan menggunakan library python, yaitu twint. Untuk dapat menggunakan twint, maka perlu menginstall library twint dengan menggunakan perintah pip install twint. Kemudian library ini akan membantu proses crawling data twitter yang dibutuhkan dengan menggunakan kata kunci sesuai kebutuhan, selanjutnya data twitter yang didapat disimpan menggunakan format file.csv.

## Analisis dan Perancangan Sistem



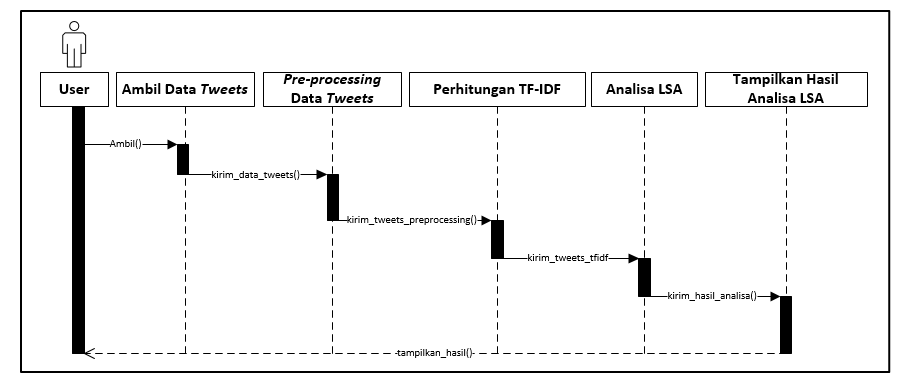
**Gambar 3.1 Blueprint Alur Sistem**

Gambar 3.1 mendeskripsikan bahwa para pengguna twitter dapat memposting *tweet*, kemudian *tweet* akan tersimpan dalam server twitter. Analis dapat mengambil data *tweets* yang tersimpan dalam server twitter menggunakan bantuan twint. Setelah *tweets* diambil, kemudian dibersihkan (*preprocessing*) untuk selanjutnya dapat dianalisa dengan menggunakan pemodelan topik *latent semantic analysis,* sehingga didapatkan topik tersembunyi dan trending topiknya. Setelah selesai akan ditampilkan hasil dari analisa tersebut.



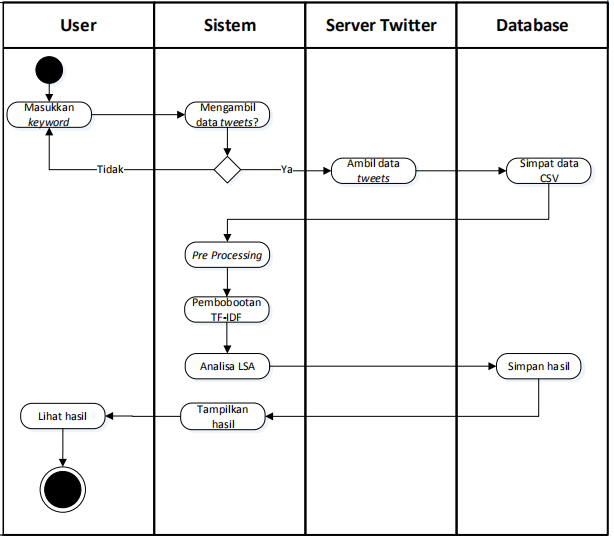
**Gambar 3.2 Use Case Diagram**

Gambar 3.2 menjelaskan bahwa Actor (User) melakukan pengambilan data tweets, dilanjutkan dengan melakukan pembersihan data dengan melakukan *pre-processing tweets,* setelah data *tweets* bersih maka dapat dilakukan perhitungan bobot *term* (istilah/kata) data *tweets* dengan melakukan perhitungan TF-IDF. Setelah perhitungan bobot *term* selesai, kemudian dilakukan analisa LSA dengan menggunakan *Singular Value Decomposition* sehingga didapatkan topik *tweets* yang tersembunyi. Tampilkan sebagai hasil.



**Gambar 3.3 Sequence Diagram**

Gambar 3.3 menerangkan bahwa user mengambil data tweets (disimpan dalam bentuk file.csv), kemudian dilakukan *pre-processing*, setelah itu dilakukan perhitungan bobot term dengan menggunakan TF-IDF. Setelah perhitungan bobot *term* selesai, kemudian dilakukan analisa LSA dengan menggunakan *Singular Value Decomposition* sehingga didapatkan topik *tweets* yang tersembunyi. Tampilkan sebagai hasil.



**Gambar 3. 4 Activity Diagram**

Dari gambar 3.4 dapat diketahui terdapat 4 bagian yang ada diantaranya User, Sistem, Server Twitter dan Database Pada setiap bagian memiliki kegiatan yang saling berhubungan. User menginputkan *keyword* kemudian sistem mengambil data tweet dari server twitter. Setelah data didapatkan sistem melakukan *pre-processing*, perhitungan bobot *term* dengan menggunakan TF-IDF, analisa LSA dengan menggunakan perhitungan *Singular Value Docomposition.* Setelah didapatkan topik tersembunyi dalam *tweets* maka dapat ditampilkan sebagai hasil.

**Gambar 3. 5 Class Diagram**

### Kebutuhan Proses

Secara umum terdapat 3 proses dalam penelitian ini, diantaranya :

Proses Input

API Twitter yang digunakan untuk *crawling tweets*

Tweets yang akan dianalisis menggunakan latent semantic analysis

Masukkan kata kunci @ganjarpranowo pada script untuk mendapat

tweets yang nantinya akan dilakukan proses clustering berdasarkan kesamaan topik.

Proses Clustering

Masukkan kata kunci nama akun twitter

Sistem akan melakukan crawling tweets yang berkaitan dengan kata kunci.

Setelah diperolah tweets, maka dilakukan tahap preprocessing yang terdiri dari :

* 1. *Case Folding*

*Case Folding* adalah proses mengubah semua huruf dalam tweet menjadi huruf kecil (lowercase). Hanya huruf “a” sampai dengan “z” yang diterima dan karakter selain huruf dihilangkan.

* 1. *Tokenizing*

*Tokenizing* adalah proses pemotongan kalimat (*string*) berdasarkan kata peyusunnya.

* 1. *Filtering*

*Filtering* adalah tahapan mengambil kata - kata penting dari hasil token dan membuang sebagian kata tertentu.

Cluster data sesuai dengan acuan yang telah dibuat, yaitu berdasarkan kemiripan topik.

Setelah ditemukan topik, tampilkan dalam bentuk grafik.

Proses output

Hasil clustering data berupa trending topik yang berasal dari tweets dengan kata kunci ganjar pranowo.

## Analisis dan Rancangan Sistem

Untuk membangun sistem yang dapat melakukan clustering, diperlukan *blueprint* sitem utama yang jelas, sehingga dibuat *blueprint* sistem yang diilustrasikan pada gambar 3.1.

**Gambar 3.1**

Pengguna twitter memposting tweet yang berhubungan dengan pejabat publik, dalam penelitian ini adalah @ganjarpranowo selaku gubernur jawa tengah, Indonesia.

## Prosedur dan Pengumpulan Data

DAFTAR PUSTAKA

1. Latar Belakang

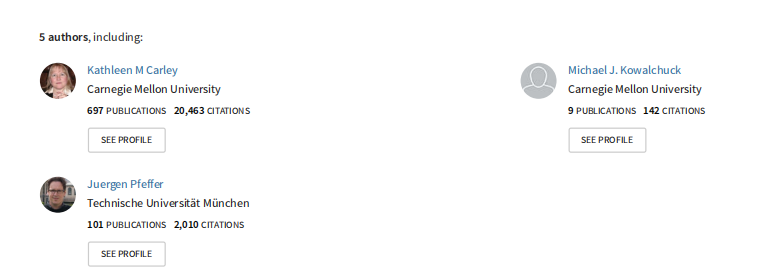
Matwin S. 2013. Institue for big data analytics: Message from the director [internet].

[diakses 2020 Januari 13]. Tersedia pada: <https://bigdata.cs.dal.ca/about>.

Singh, S. (2015): Spatial Temporal Analysis of Social Media Data. Master Thesis, Technische Universitat Munchen.

Carley KM, Malik M, Kowalchuk M, Pfeffer J dan Landwehr P. 2015. *Twitter usage in indonesia*. USA: Center for the Computational Analysis of Social and Organizational Systems.

Nair G. 2016. Text Mining 101: Topic Modeling [internet]. [diakses 2020 Februari 1].



1. Teori

Doig C. 2015. Introduction to topic modeling in python [internet]. [diakses 2017

Mei 10]. Tersedia pada: http://chdoig.github.io/pytexas2015-topic-modeling.

Tersedia pada: http://www.kdnuggets.com/2016/07/text-mining-101-topic

modeling.html.

<https://pypi.org/project/twint/> ini adalah twint akses mei

# Text Preprocessing BY [INFORMATIKALOGI](https://informatikalogi.com/author/cskaditya/" \o "Posts by informatikalogi) · PUBLISHED NOVEMBER 27, 2016 · UPDATED JULY 11, 2017 <https://informatikalogi.com/text-preprocessing/>

<https://medium.com/@ksnugroho/dasar-text-preprocessing-dengan-python-a4fa52608ffe> pre processing juga itu

Bagaimana menemukan topic tersembunyi dari tweets yang ada menggunakan latent semantic analysis dengan menggunakan bahasa pemrograman python.

Pemodelan topik akan mengidentifikasi topic yang disajikan dalam dokumen, sementara klasifikasi teks mengklasifikasikan teks ke dalam suatu kelas.