UNIVERSITAS
MUHAMMADIYAH
SIDOARJO

UMSIDA
DARI SINI PENCERAHAN BERSEMI

Fitur Ekstraksi Pada Pemodelan Topik Menggunakan Metode Latent Dirichlet Allocation Pada Peristiwa Kebocoran Data

Oleh:

Achmad Ariansyah

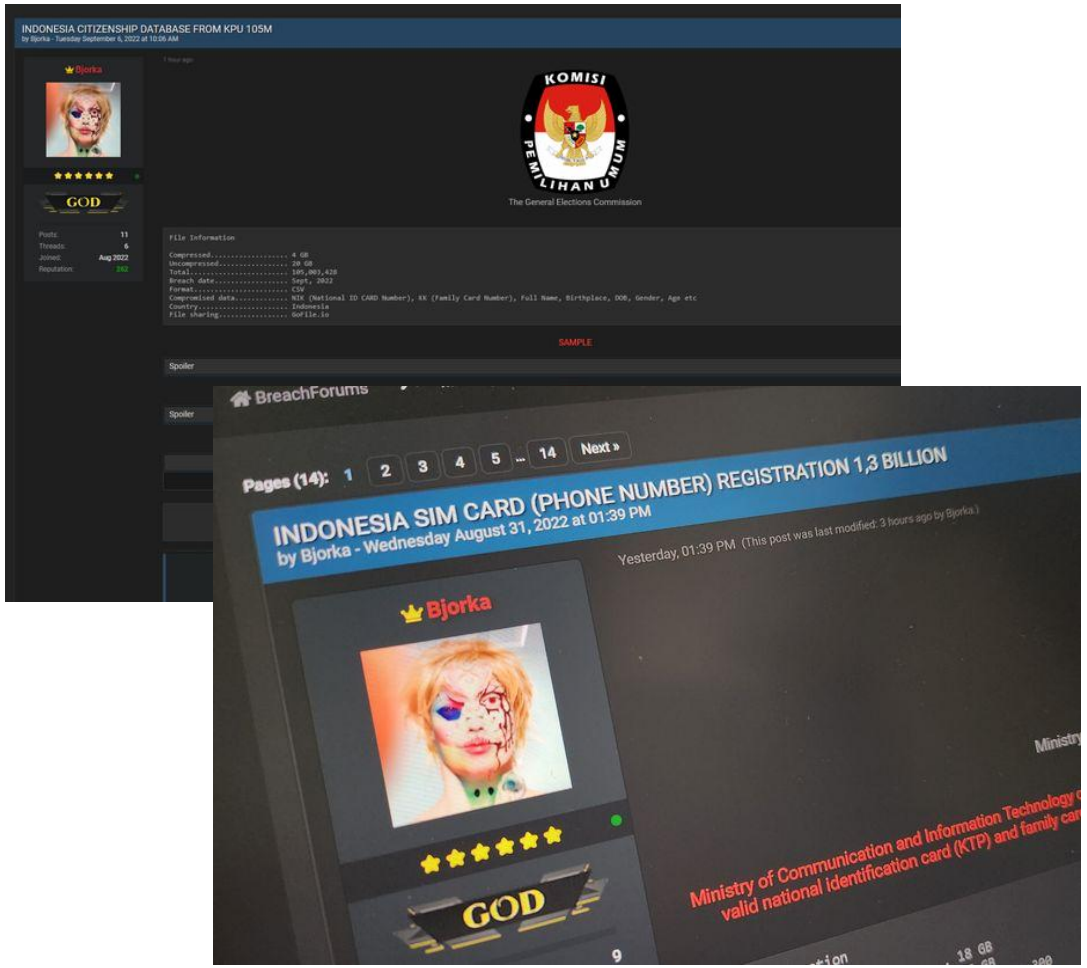
Program Studi

Informatika

Universitas Muhammadiyah Sidoarjo

2023

Latar Belakang



Peristiwa kebocoran data yang terjadi pada bulan September 2022 sangat merugikan banyak pihak.

Data yang bocor diantaranya yaitu data registrasi sim card, data KPU RI penduduk sebanyak 105jt, data pelanggan provider indihome.

Latar Belakang

Peristiwa tersebut memancing masyarakat untuk menuliskan opini mereka dengan bebas di aplikasi Twitter.

Dengan banyaknya opini tersebut maka bisa dibuatkan pemodelan topik untuk mengelompokkan berbagai opini dari masyarakat

Rumusan Masalah

1. Bagaimana cara memodelkan topik mengenai peristiwa kebocoran data pada media sosial twitter menggunakan metode LDA?
2. Bagaimana perbandingan menggunakan fitur ekstarksi tf-idf dan bag of word?

Penelitian Terdahulu

Herwanto, 2018

Judul : Document Clustering Dengan Latent Dirichlet Allocation dan Ward Hierarichal Clustering

Published : Jurnal Pseudocode (Universitas Brawijaya)

- Penelitian ini membahas tentang document clustering dari konten informasi dalam bentuk berita menggunakan metode LDA dan Ward Hierarichal Clustering dan mempertimbangkan nilai silhouette coefficient untuk menentukan jumlah topic

Penelitian Terdahulu

Wirasakti dkk., 2020

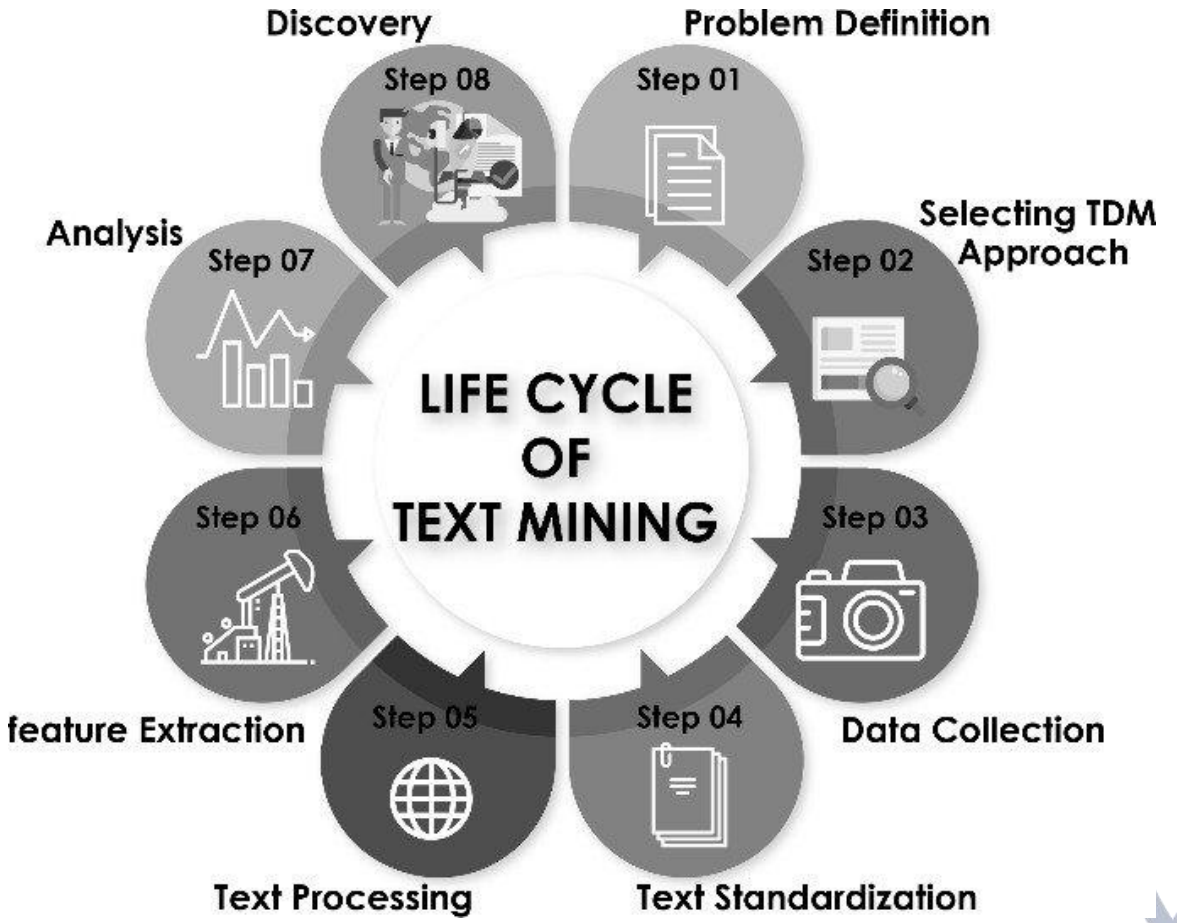
Judul : Pembuatan Kata Kunci Otomatis Dalam Artikel Dengan Pemodelan Topik

Published : Jurnal Media Informatika Budidarma

- Penelitian ini membahas tentang pembuatan kata kunci otomatis untuk publikasi artikel pada sebuah blog menggunakan metode LDA dan pengelompokan menggunakan K-Means untuk membantu mencari topik dengan nilai yang tinggi.

Metode Penelitian

Metode Penelitian ini menggunakan *Life Cycle Of Text Mining* yang dikemukakan oleh Shaikh dkk., 2019



Metode

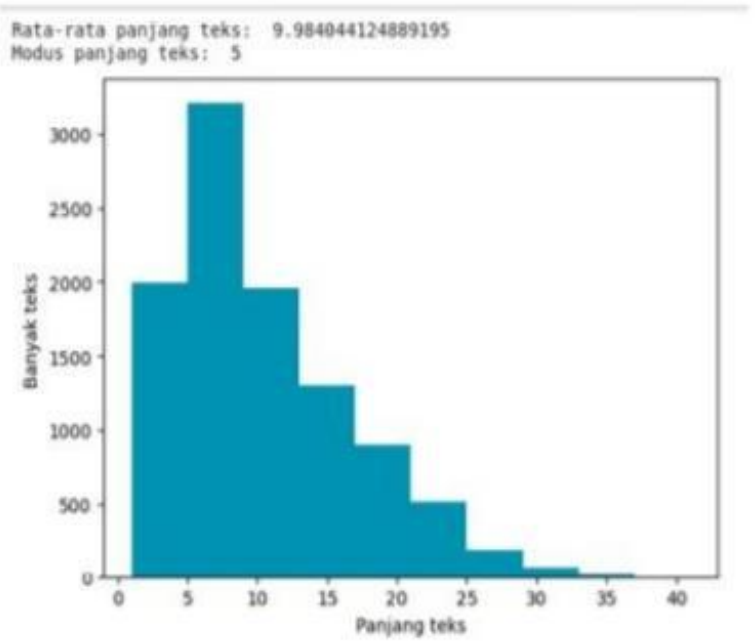
Metode yang akan digunakan pada penelitian ini yaitu :

1. TF-IDF dan Bag of Word untuk melakukan ekstraksi fitur data mentah
2. Latent Dirichlet Allocation untuk melakukan Pemodelan Topik

Hasil

Hasil dari Preprocessing Data :

	word	freq
103	pribadi	1546
12	kominfo	1195
307	bjorka	1065
206	negara	981
109	rakyat	975
52	sim	939
152	perintah	929
18	nik	870
37	bocor	810
55	aman	769



Hasil

- Pada tahap ini dilakukan pemodelan topik LDA, melakukan iterasi sebanyak 40 kali dan memberi batasan pada 5 topik untuk menentukan score coherences.

Jumlah Topik	BoW	TF-IDF
1	0.4589483707911217	0.29970908185021 28
2	0.4379672499472 2813	0.3412056563073 3674
3	0.472835505696 4541	0.2277983754809 6019
4	0.4132882189811 5087	0.4162606079798459
5	0.4100343095293712	0.4797712177878327

Pembahasan

Jumlah Topik	BoW
1	'orang', 'rakyat', 'perintah', 'nik', 'pribadi', 'kominfo', 'jaga', 'masyarakat', 'data', 'kerja'
2	'johnny', 'bjorka', 'plate', 'pribadi', 'bocor', 'hacker', 'menkominfo', 'indonesia', 'negara', 'sim'
3	'lindung', 'negara', 'pdp', 'pribadi', 'uu', 'bbm', 'rakyat', 'ruu', 'sistem', 'digital'

1. Menginterpretasikan topik tentang kominfo harus menjaga data pribadi seperti NIK rakyat

2. Menginterpretasikan topik tentang johnny g plate selaku menkominfo di indonesia bertanggung jawab atas kasus kebocoran data ulah hacker bjorka.

3. Menginterpretasikan topik tentang perlindungan data pribadi rakyat memalui ruu pdp.

Pembahasan

Jumlah Topik	TF-IDF
1	'negara', 'pribadi', 'bjorka', 'rakyat', 'kominfo', 'lindung', 'hacker', 'indonesia', 'aman', 'duga'
2	'mahfud', 'md', 'pdp', 'tegas', 'bin', 'polri'. 'apaapanya', 'sah', 'serang', 'isu'
3	'orang', 'ri', 'gin', 'ktp', 'kerja', 'ganti', 'tanggung', 'presiden', 'pake', 'suruh'
4	'cek', 'anggota', 'pilih', 'mudah', 'heran', 'bri', 'ngurusin', 'hoax', 'usang', 'lho'
5	'tim', 'lemah', 'an', 'viral', 'mpud', 'pegawai', 'pantes', 'darkweb', 'khusus', 'lawak'

Topik pertama menginterpretasikan topik tentang kominfo lindungi data pribadi dari hacker bjorka.

Pada Topik kedua isi dari topik kedua ini susah untuk diintrepretasikan karena kata kata yang dihasilkan tidak saling berhubungan.

Untuk topik ketiga sampai topik kelima juga susah untuk diinterpretasikan.

Kesimpulan

- Pada penelitian ini diaplikasikan fitur ekstraksi Bag of Word dan TF-IDF pada pemodelan topik Latent Diriclet Allocation terhadal 11.067 data
- Dengan perbandingan fitur ekstraksi BoW dan TF-ID didapatkan hasil fitur ekstraksi BoW lebih baik daripada fitur ekstraksi TF-IDF


UMSIDA
DARI SINI PENCERAHAN BERSEMI