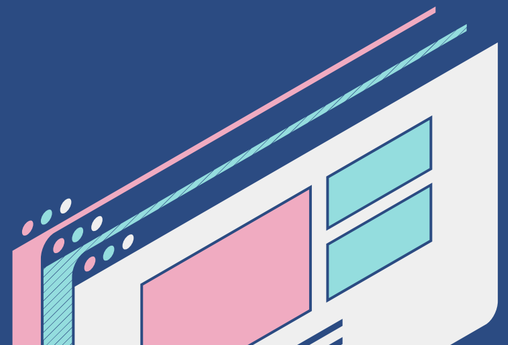
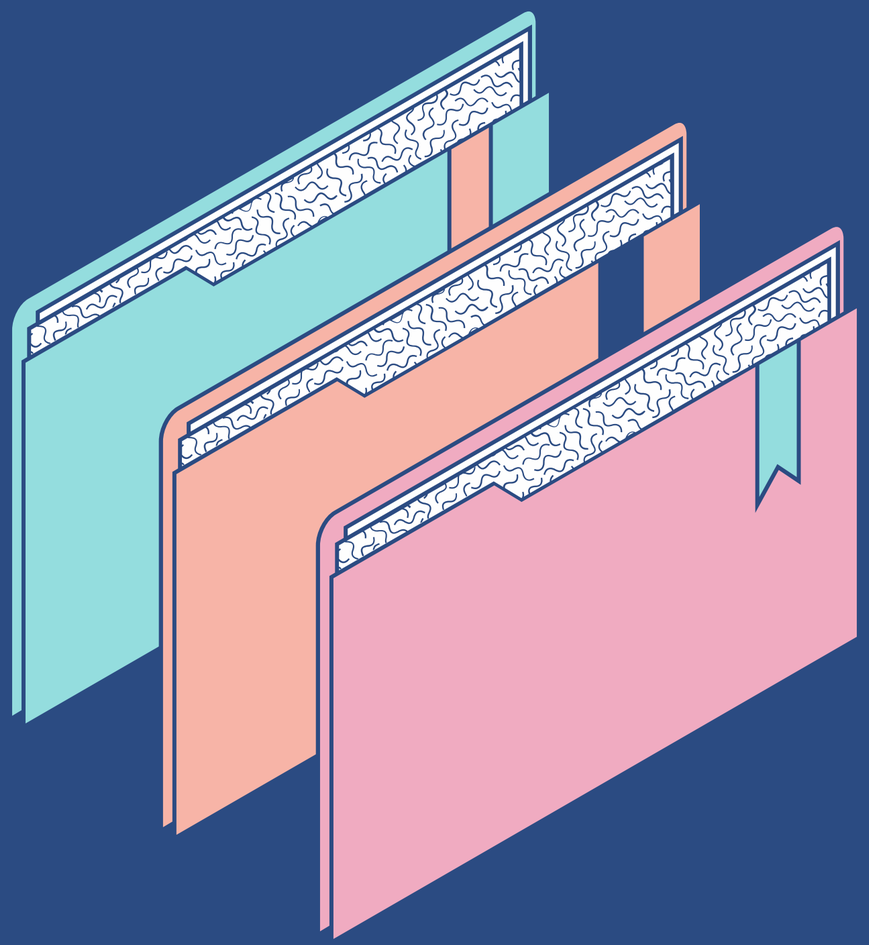


SEMINAR HASIL TUGAS AKHIR

Analisis Sentimen Mengenai Undang - Undang TPKS pada Media Sosial Twitter Menggunakan Metode Support Vector Machine dan K-Nearest Neighbour

Arviandri Naufal Zaki - 064001800035

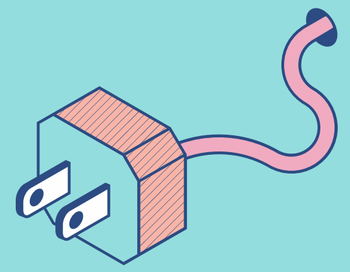


Latar Belakang

Twitter oleh masyarakat Indonesia dimanfaatkan untuk berbagai hal seperti berkomunikasi dengan orang lain secara publik atau personal, berbagi kabar dan opini pribadi, berjualan, sampai mengkritik atau memuji akan suatu hal.

Pemerintah juga memanfaatkan platform ini untuk mengetahui respon masyarakat kepada kebijakan yang baru dikeluarkan. Oleh karena itu, pengguna Twitter dapat beropini yang dipengaruhi oleh emosi yang dapat diklasifikasikan untuk menentukan polarisasinya.

Rumusan Masalah



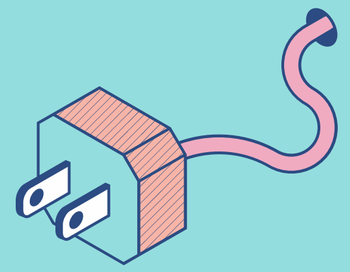
Bagaimana cara mengambil dan mengolah data tweet yang berasal dari Twitter untuk perhitungan Support Vector Machine (SVM) dan K-Nearest Neighbour (KNN)

Bagaimana tingkat keakuratan dari K-Nearest Neighbour (KNN) dan Support Vector Machine (SVM) pada analisis sentimen di Twitter mengenai Undang – Undang TPKS.

Bagaimana hasil klasifikasi dari tweet menggunakan Support Vector Machine (SVM) dan K-Nearest Neighbour (KNN) pada analisis sentimen di Twitter mengenai Undang – Undang TPKS.

Bagaimana cara implementasi sentimen analisis secara hybrid menggunakan lexicon based dan SVM serta KNN.

Batasan Masalah



Data yang digunakan adalah tweet berbahasa Indonesia dengan kata kunci “uu tpks” dari Twitter.

Metode yang digunakan untuk klasifikasi adalah Support Vector Machine (SVM) dan K-Nearest Neighbour (KNN).

Metode yang digunakan untuk pelabelan data adalah Valence Aware Dictionary and sEntiment Reasoner (VADER).

Data yang digunakan adalah tweet berbahasa Indonesia dengan kata kunci “uu tpks” dari Twitter.

Tujuan Penelitian

Tujuan dari tugas akhir ini yaitu untuk mengklasifikasi tweet berdasarkan positif dan negatifnya untuk mengetahui keakuratan dari kedua metode ini yaitu Support Vector Machine (SVM) dan K-Nearest Neighbour (KNN) dalam menganalisis sentimen (emosi) pengguna Twitter mengenai Undang – Undang TPKS.

Manfaat Penelitian

Memperoleh visualisasi sentimen analisis berupa feedback dari pengguna twitter mengenai Undang – Undang TPKS dengan menggunakan metode SVM dan KNN.

Bagi pemerintah dapat mengetahui sentimen yang didapatkan dari pengesahan UU TPKS dan dapat digunakan sebagai rujukan untuk memperbaharui kebijakan lain yang dikeluarkan.

Memperoleh perbandingan akurasi dari metode SVM dan KNN pada penelitian ini.

Memperoleh perbandingan akurasi dari penggunaan linear dan RBF untuk metode SVM pada penelitian ini.

Penelitian Sebelumnya

No	Nama Penulis - Tahun	Judul	Hasil Kajian
1	Nabilah Putri Aprilia, Dian Pratiwi, dan Anung Barlianto Ariwibowo - 2019	Sentiment Visualization Of Covid-19 Vaccine Based On Naïve Bayes Analysis	Pada penelitian ini peneliti melakukan panggilan API ke twitter untuk mendapatkan data, kemudiandari hasil dari data yang diambil tersebut dilakukan preprosesing pada data tersebut dan dilanjutkan dengan labeling untuk menentukan positif dan negatifnya lalu dilakukan ekstraksi data menggunakan TF-IDF untuk pembobotan kata, lalu dilakukan perhitungan menggunakan Naïve Bayes dan mendapatkan nilai akurasi 73,1% ,presisi 73 %, dan recall 83%
2	Ghulam Asrofi Buntoro - 2017	Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter	Dari penelitian ini didapatkan hasil yang cukup tinggi untuk klasifikasi menggunakan Naive Bayes dibandingkan menggunakan SVM yaitu sekitar 95% nilai akurasi, 95% nilai presisi, dan 95% nilai recall.
3	Gary Dimitri Hamidi, Farida Afira Bestari, dan Alexandra Situmorang - 2021	Sentiment Analysis on the Ratification of Penghapusan Kekerasan Seksual Bill on Twitter	Pada penelitian ini peneliti menggunakan beberapa klasifikasi yaitu SVM, Bernoulli, dan Logistic Regression yang masing – masing menghasilkan keakuratan hingga 63 %, 65 %, dan 65 % serta peneliti menyimpulkan bahwa kata “kekerasan”, “korban” dan “seksual” pada topik ini mengekspresikan sentimen positif
4	T Mustaqim, K Umam, dan M A Muslim - 2020	Twitter text mining for sentiment analysis on government's response to forest fires with vader lexicon polarity detection and k-nearest neighbor algorithm	Lalu di penelitian ini, peneliti menggunakan cara penggabungan antara metode Vader dengan KNN dan menghasilkan hasil akurasi yang cukup baik yaitu 75 %

Landasan Teori

Twitter

Twitter adalah platform sosial media yang dapat digunakan untuk mengirimkan suatu postingan (tweet) dalam bentuk foto maupun teks dengan terbatas yaitu 280 karakter

Python

Python adalah bahasa pemrograman dengan kode sumber yang terbuka (open source) yang dapat digunakan untuk membuat program secara independent (standalone) maupun untuk membuat program scripting.

Scraping Data

Teknik Scraping menggunakan cara mengambil data dari apa yang ditampilkan oleh website. Pada tahap ini dilakukan penarikan data menggunakan library snscrape.

Preprocessing

Tujuan dilakukannya preprocessing dokumen adalah untuk menghilangkan suatu hal yang dapat menggangu jalannya analisis, menyeragamkan bentuk kata dan mengurangi volume kata.

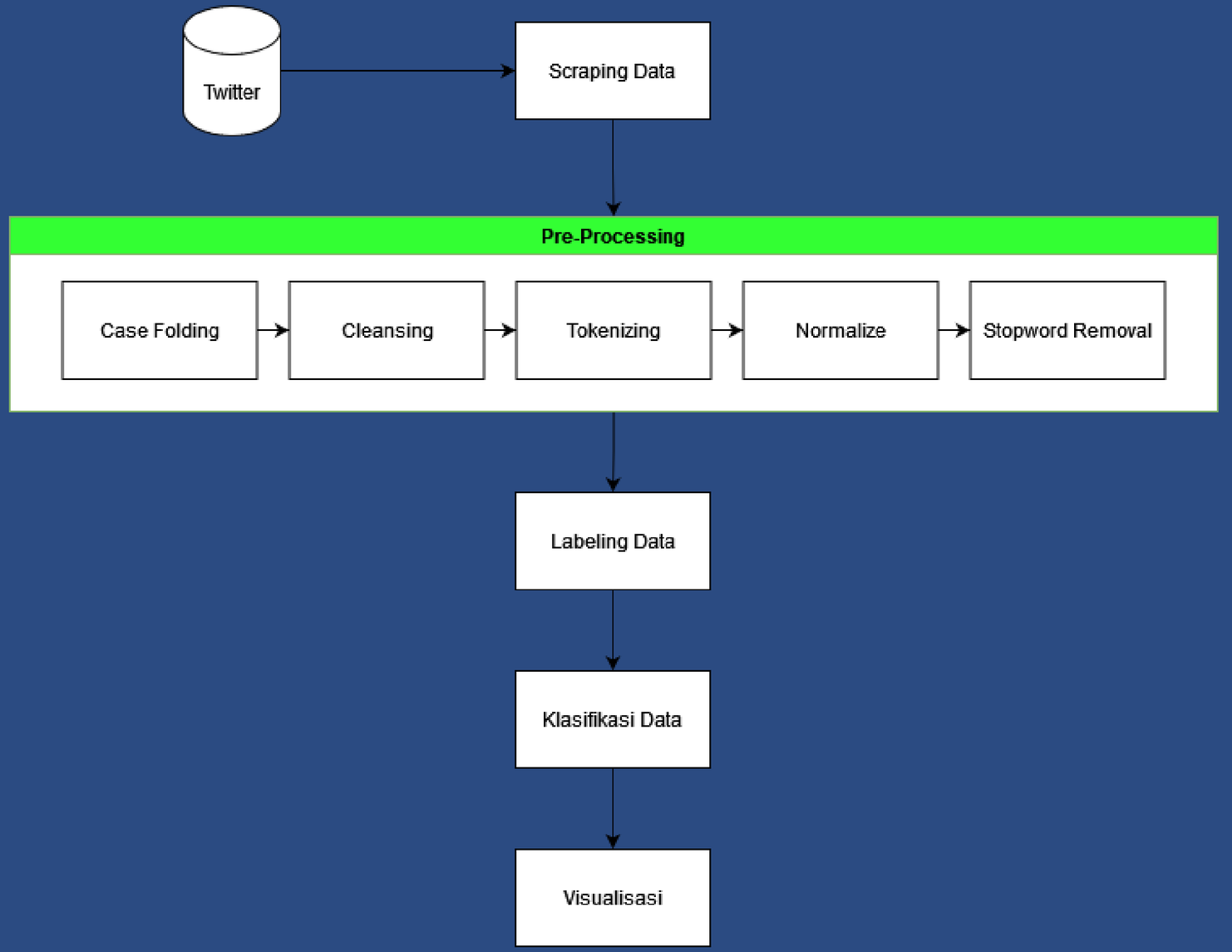
Lexicon

Lexicon merupakan kumpulan kata pada sentimen yang telah diketahui dan dihimpun dalam bentuk dataset

TF-IDF

TF-IDF merupakan suatu algoritma yang dapat menghasilkan informasi tentang seberapa sering kata tersebut muncul di dalam dataset tersebut dan dimunculkan dalam bentuk berat per kata.

Metode Penelitian



Pembahasan

Pengumpulan Data

Pengumpulan data pada penelitian ini menggunakan library snscrape. Snscrape adalah suatu library yang berisi beberapa fungsi yang dapat digunakan untuk menarik data dari sosial media seperti facebook, Instagram, twitter, dan seterusnya.

Lalu query search yang digunakan pada penelitian ini yaitu “UU TPKS” untuk kata kunci, 12 - 24 April 2022 untuk tanggal awal dan akhir, serta Indonesia (id) untuk Bahasa.

Dari pengumpulan data tersebut dihasilkan 15.632 Tweet

Pembahasan

Pengolahan Data (Pre-Processing)

Pada tahap ini dilakukannya pembersihan dan pengubahan terhadap data yang telah dihimpun sebelumnya agar tidak terdapat data yang dapat menggangu jalannya analisis dan untuk menjadikan data dapat diproses ke tahap selanjutnya.

Pada tahap ini terdapat 5 tahap yang digunakan untuk menjadikan data bersih dan siap untuk digunakan untuk tahap selanjutnya yaitu :

- Case Folding
- Cleaning
- Tokenizing
- Normalize
- Stopword Removal

Pembahasan - Pre-Processing

Case Folding

Tahap ini bertujuan untuk merubah huruf kapital menjadi huruf kecil agar datanya sama rata.

Sebelum "*Case Folding*"	Setelah "*Case Folding*"
Lihat tanggal chatnya, kalau setelah April udah bisa dijerat UU TPKS	lihat tanggal chatnya, kalau setelah april udah bisa dijerat uu tpks

Cleansing

Tahap ini bertujuan untuk menghilangkan data hashtag, mention, tanda baca, angka, url, space yang tidak berguna, serta data yang bukan ASCII seperti emotikon, data berbahasa china, dan seterusnya

Sebelum "*Cleaning*"	Setelah "*Cleaning*"
lihat tanggal chatnya, kalau setelah april udah bisa dijerat uu tpks 😅 https://t.co/c89NcYzjQv	lihat tanggal chatnya, kalau setelah april udah bisa dijerat uu tpks
UUTPKS diterapkan keras mulai dari pemrentah dan @DPR_RI, setuju? @KemensetnegRI https://t.co/isAu9nBZpx	UUTPKS diterapkan keras mulai dari pemrentah dan setuju

Pembahasan - Pre-Processing

Tokenizing

Di tahap ini data akan dipisahkan bedasarkan separator (space) menjadi token – token (kata di setiap kalimat pada data). Langkah ini berfungsi untuk menjadikan data kompartibel dengan tahap selanjutnya.

Sebelum "*Tokenizing*"	Setelah "*Tokenizing*"
lihat tanggal chatnya, kalau setelah april udah bisa dijerat uu tpks	[lihat, tanggal, chatnya, kalau, setelah, april, udah, bisa, dijerat, uu, tpks]

Normalize

Pada tahap ini data yang berbentuk tidak baku diubah menjadi kata baku. Langkah ini berfungsi untuk mencegah terjadinya terdapat kata – kata yang diluar lexicon (out of vocabulary) dikarenakan sebagian besar lexicon adalah kata baku.

Sebelum "*Normalize*"	Setelah "*Normalize*"
kagak ada yg ribut nyuruh seragaman baju nasional/kebaya uu tpks sah vaksin serviks bakal jd vaksin wajib	tidak ada yang ribut nyuruh seragaman baju nasional kebaya uu tpks sah vaksin serviks bakal jadi vaksin wajib

Pembahasan - Pre-Processing

Stopword removal

Tahapan ini bertujuan untuk menghilangkan kata – kata yang tidak diperlukan pada data yang dapat menggangu jalannya analisis.

Sebelum "*Stopword Removal*"	Setelah "*Stopword Removal*"
lihat tanggal chatnya kalau setelah april sudah bisa dijerat uu tpks	tanggal chatnya april dijerat uu tpks

Pembahasan

Labeling Data

Tahap ini bertujuan untuk mengklasifikasi data menjadi sentiment (positif & negatif). Klasifikasi yang digunakan berbasis Lexicon – based dengan menggunakan library VADER (Valence Aware Dictionary and sEntiment Reasoner). Kamus yang terdapat pada Lexicon berjumlah 3610 kata positif dan 6670 kata negatif.

	tweet	sentimen
2	menangani kekerasan seksual disahkan enam dibahas dewan perwakilan rakyat mengesa	Negatif
3	menangani kekerasan seksual disahkan enam dibahas dewan perwakilan rakyat mengesa	Negatif
4	wakil ketua mpr ri mahasiswa kawal implementasi	Negatif
5	fadel muhammad mahasiswa kawal implementasi	Negatif
6	tanggal chatnya april dijerat	Negatif
7	kelakuan gini ranah	Netral
8	membuka sistem peradilan diharapkan implementasinya ks diselesaikan adil korban pela	Negatif
9	mochammad abizar yusro terang sinar perlindungan kartini	Negatif
10	puan pengesahan undangundang bentuk hadiah perempuan indonesia menjelang kartini	Positif

Pada tahap ini dari 15.632 data menghasilkan label 7244 (46%) data positif, 5416 (35%) data negatif, 2972 (19%) data netral.

Pembahasan

Pembobotan Kata (Feature Extraction)

Di tahap ini dilakukan proses untuk mengubah kata – kata yang terkumpul menjadi vektor agar data dapat digunakan di proses selanjutnya. Metode dari pembobotan kata yang digunakan yaitu TF-IDF. Rumus dari TF-IDF & IDF adalah sebagai berikut:

$$TF - IDF_{t,d} = TF_{t,d} \times IDF_t$$

$TF_{t,d}$: Frekuensi kata terhadap kata t di dokumen d

IDF_t : Kejarangan frekuensi kata t pada dokumen

$$IDF_t = \ln\left(\frac{1+N}{1+df_t}\right) + 1$$

Keterangan :

N : Jumlah dokumen.

df_t : Jumlah dokumen yang terdapat kata t.

TF-IDF yang terdapat pada library scikit-learn vectornya juga dinormalisasikan menggunakan rumus equlidian. Rumusnya adalah sebagai berikut :

$$v_{norm} = \frac{v}{\sqrt{v_1^2 + v_2^2 + v_3^2 + v_4^2 + \cdots + v_n^2}}$$

v_{norm} : Vektor TF-IDF setelah normalisasi.

v : Vektor TF-IDF sebelum normalisasi.

$v_1^2 + v_2^2 + v_3^2 + v_4^2 + \cdots + v_n^2$: Vektor yang terdapat pada dokumen yang sama.

Pembahasan

Klasifikasi Data

Sebelum data diklasifikasikan, data yang berlabel netral dijadikan sebagai positif dikarenakan SVM hanya menerima label binary. Lalu setelah itu data dibagi terlebih dahulu menjadi data latih (train) dan data uji (test) dengan masing - masing sebanyak 90% dan 10% dari 15.632 data.
Lalu data diklasifikasikan menggunakan 2 metode yaitu Support Vector Machine dan K-Nearest Neigbour.

Support Vector Machine adalah metode klasifikasi yang menggunakan cara mengklasifikasikan secara linear dengan menemukan hyperlane yang terbaik yang berfungsi sebagai pemisah antara 2 kelas.

Sedangkan K-Nearest Neighbour adalah sebuah algoritma untuk klasifikasi yang menggunakan cara mengukur tingkat kemiripan antar data yang bertetangga (cosine similarity) atau mengukur jarak euclidean dari data latih (training data) dengan data uji (test data)

Pembahasan

Hasil Akurasi dari Klasifikasi Data

	Kelas	Prediksi	
		Negatif	Positif
Aktual	Negatif	1256	26
	Positif	96	186

SVM Kernel Linear

	Kelas	Prediksi	
		Negatif	Positif
Aktual	Negatif	1272	10
	Positif	119	163

SVM Kernel RBF

	Kelas	Prediksi	
		Negatif	Positif
Aktual	Negatif	1256	26
	Positif	96	186

KNN dengan K = 4

Pembahasan

Hasil Akurasi dari Klasifikasi Data

Klasifikasi	Akurasi
SVM Kernel Linear	0.921995 (92.2%)
SVM Kernel RBF	0.917519 (91.8%)
KNN dengan K = 4	0.753197 (75.3%)

Pembahasan

Hasil Visualisasi

Positif

Netral

Negatif

Terima Kasih