UNIVERSITAS TRISAKTI

Analisis Sentimen Terhadap Penurunan Harga Telur pada Data Twitter Menggunakan Metode Support Vector Machine dan K-Nearest Neighbour

Diajukan Sebagai Syarat Dalam Memperoleh Gelar Sarjana Strata Satu (S1)
Program Studi Sistem Informasi

PROPOSAL TUGAS AKHIR

Disusun Oleh:
Nama : Arviandri Naufal Zaki
NIM : 064001800035

FAKULTAS TEKNOLOGI INDUSTRI

PRODI SISTEM INFORMASI

UNIVERSITAS TRISAKTI

JAKARTA BARAT

HALAMAN PENGESAHAN

Analisis Sentimen Terhadap Penurunan Harga Telur pada Data Twitter Menggunakan Metode Support Vector Machine dan K-Nearest Neighbour

PROPOSAL TUGAS AKHIR

Diajukan Sebagai Syarat Dalam Memperoleh Gelar Sarjana Strata Satu (S1)
Program Studi Sistem Informasi

Universitas Trisakti

Disusun Oleh:
Nama : Arviandri Naufal Zaki
NIM : 064001800035

Jakarta, 23 Oktober 2021

Pembimbing Utama	Pembimbing Pendamping
Dian Pratiwi, ST, MTI	Anung B. Ariwibowo, M.Kom

ABSTRAK

Nama : Arviandri Naufal Zaki

Program Studi : Teknik Informatika

Judul : Analisis Sentimen terhadap Penurunan Harga Telur pada data Twitter Menggunakan Metode KNN dan SVM

Twitter adalah media sosial yang banyak digunakan oleh masyarakat Indonesia maupun Dunia. Twitter juga dimanfaatkan untuk berbagi kabar dan opini pribadi, memasarkan produk, sampai mengkritik suatu kebijakan atau peraturan. Opini yang diposting sebagai tweet di Twitter juga dapat digunakan sebagai tolak ukur apakah kebijakan yang dikeluarkan banyak yang mendukungnya atau sebaliknya. Untuk memperoleh tolak ukur tersebut maka digunakanlah analisis sentimen untuk memisahkan opini positif dengan opini negatif. Dari pengambilan data untuk diproses maka digunakanlah *scraping* dari website Twitter untuk mendapatkannya. Setelah itu dilakukan proses awal sebelum data diolah yaitu *Preprocessing* untuk menghilangkan bagian yang tidak berguna dalam pengolahan data. Lalu dilakukan teknik Support Vector Machine dan K-Nearest Neighbour untuk mengklasifikasikan opini postif dan negatif guna untuk membandingkan manakah yang lebih banyak dari opini tersebut lalu dijadikanlah tolak ukur terhadap suatu kebijakan.

Kata Kunci : *Analisis Sentimen, Twitter, Penurunan Harga Telur, Support Vector Machine, Scraping*

1. Latar Belakang

Perkembangan teknologi berkembang pesat termasuk didalamnya yaitu komunikasi, media sosial adalah salah satunya. Media sosial juga merupakan salah satu layanan internet yang sering dipakai oleh masyarakat di Indonesia. Salah satu dari media sosial tersebut adalah Twitter.

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 suatu hal. Dikarenakan informasi yang berada di Twitter juga dibatasi sekitar 280 karakter biasanya pengguna hanya mengirim suatu hal yang pendek [1]. Pemerintah juga memanfaatkan *platform* ini untuk menginformasikan tentang perubahan harga kebutuhan pokok. Oleh karena itu, pengguna Twitter dapat beropini yang dipengaruhi oleh emosi yang dapat diklasifikasikan untuk menentukan polarisasinya, yaitu positif atau negatif tentang *tweet* oleh pemerintah yang telah disebutkan sebelumnya.

Analisis sentimen yaitu kegiatan mengolah kata untuk menghasilkan suatu sentimen (positif atau negatif), Analisis sentimen bertujuan salah satunya yaitu untuk mendapatkan suatu opini dari kebijakan pemerintah yang telah dikeluarkan kemudian opini tersebut diklasifikasikan ke dalam sentimen positif dan negatif. Teknik yang dipakai untuk mengambil data dari Twitter sebelum di analisis yaitu menggunakan teknik *Scraping* yaitu mengambil data langsung dari website Twitter. Lalu teknik yang digunakan untuk mengklasifikasi data tersebut yaitu *K-Nearest Neighbor (KNN)*, *Support Vector Machine (SVM)*, dan *Naïve Bayes*.

2. Rumusan Masalah

Berdasarkan latar belakang yang telah disusun sebelumnya, rumusan masalahnya yaitu :

a. Bagaimana cara mengambil dan mengolah data tweet yang berasal dari Twitter untuk perhitungan *Support Vector Machine* (SVM) dan *K-Nearest Neighbour* (KNN).
b. Bagaimana tingkat keakuratan dari *K-Nearest Neighbour* (KNN) dan *Support Vector Machine* (SVM) pada analisis sentimen di Twitter mengenai penurunan harga telur.
c. Bagaimana hasil klasifikasi dari tweet menggunakan *Support Vector Machine* (SVM) dan *K-Nearest Neighbour* (KNN).

3. Batasan Masalah

Berdasarkan latar belakang yang telah disusun sebelumnya, rumusan masalahnya yaitu :

a. Data yang digunakan adalah tweet berbahasa Indonesia dengan kata kunci "Harga Telur Turun" dan "Harga Telor Turun" dari Twitter.
b. Metode yang digunakan untuk klasifikasi adalah *Support Vector Machine* (SVM) dan *K-Nearest Neighbour* (KNN).

4. 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 penurunan harga telur.

5. Manfaat Penelitian

Manfaat penulisan tugas akhir ini adalah:

a. Memperoleh hasil analisis sentimen terhadap penurunan harga telur dengan menggunakan metode SVM dan KNN.
b. Bagi pemerintah dapat digunakan untuk meningkatkan pengetahuan dan dapat digunakan sebagai rujukan untuk memperbaharui kebijakan yang dikeluarkan.

6. Kajian Pustaka

6.1. Penelitian Terdahulu

Melakukan penulisan untuk penelitian membutuhkan sebuah panduan dan dukungan dari penelitian yang sudah terlebih dahulu ada sebelumnya yang juga berkaitan dengan penelitian yang sedang berlangsung.

Pada penelitian yang telah dilakukan sebelumnya dapat disimpulkan bahwa penelitiannya memiliki tantangan terbesar salah satunya dalam melakukan pengambilan data dari Twitter masih menggunakan API yang diberikan oleh Twitter. Dengan menggunakan cara tersebut maka data tweet yang didapatkan hanya dalam batas waktu seminggu ke belakang dari hari ini dan pengambilan data setiap hari dibatasi hanya 50.000 tweet per hari [1].

Kemudian dari 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* sehingga dapat dilakukannya penentuan sentiment (positif atau negatif) pada penelitian tersebut [2] .

Lalu bedasarkan dari penelitian sebelumnya dapat diketahui bahwa akurasi dari metode *Naïve Bayes* 95 % sentimen cenderung negatif tetapi ketika memakai metode *Support Vector Machine* akurasinya 90% sentiment cenderung positif. Dapat disimpulkan bahwa publik memiliki perasaan baik terhadap orang tersebut [3].

6.2. Landasan Teori

6.2.1. Twitter

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

Selain daripada itu Twitter juga banyak digunakan oleh masyarakat karena kepraktisannya dalam menyampaikan sesuatu seperti mengungkapkan pendapat, berkomunikasi, dan lain sebagainya [1].

6.2.2. *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*. Python juga bahasa pemrograman yang yang dianggap paling banyak digunakan di dunia [4].

Bahasa *python* lebih mudah dipahami dikarenakan bahasa pemrograman ini lebih mendekati bahasa manusia dibandingkan bahasa pemrograman lain. Fitur yang terdapat pada *python* juga beragam seperti dapat dijalankan di hampir seluruh sistem operasi (*cross platform*), program atau *script* yang mudah dipindahkan (*portable*), dan masih banyak lagi.

6.2.3. Analisis Sentimen

Analisis sentimen, yang disebut juga dengan penambangan opini (*opinion mining*), merupakan cabang ilmu dari penambangan data yang bertujuan untuk memahami, menganalisis, mengekstrak, dan mengolah data berbentuk teks yang berupa opini terhadap entitas seperti produk, servis, organisasi, individu, dan topik tertentu [5].

6.2.4. *Scraping* Data

Scraping data adalah tahap pertama yang dilakukan untuk melakukan analisis sentimen dari opini pengguna Twitter. Teknik Scraping menggunakan cara mengambil data dari apa yang ditampilkan oleh website [6] . Pada tahap ini dilakukan penarikan data menggunakan *library* snscrape, karena *library* ini dapat menarik data yang tidak dapat dilakukan oleh API Twitter gratis yaitu lebih dari 7 hari kebelakang dan dapat menarik tweet lebih dari batas API Twitter [7].

6.2.5. Preprocessing

Tujuan dilakukannya preprocessing dokumen adalah untuk menghilangkan suatu hal yang dapat menggangu jalannya analisis, menyeragamkan bentuk kata dan mengurangi volume kata. Pada tahap preprocessing ini dilakukan proses Case Folding, Cleansing, Tokenizing, Normalization ,Stopword Removing, dan Stemming.

6.2.6. 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. Untuk menentukan berat dari per kata tersebut algoritma ini menggunakan beberapa komponen yang sesuai dengan namanya yaitu *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) [8].

Term Frequency (TF) adalah seberapa sering kata tersebut muncul dalam dataset sedangkan *Inverse Document Frequency* (IDF) adalah pengurangan dari berat setiap kata yang muncul pada dataset.

Rumus dari *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) adalah sebagai berikut :

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

Keterangan :

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

IDF_t : Kejarangan frekuensi kata t pada dokumen

$$IDF_t = \log\left(\frac{N}{df_t}\right)$$

Keterangan :

N : Jumlah dokumen.

df_t : Jumlah dokumen yang terdapat kata t.

Contoh :

Terdapat kalimat “Saya sedang belajar hitung - hitung tf idf.” tentukan nilai TF dan TF-IDF !

Jawaban ;

Tabel TF :

Term (t)	D1 (Dokumen 1)
Saya	1
sedang	1
belajar	1
hitung	2
tf	1

idf	1

Nilai DF (*Document Frequency*) :

Term (t)	DF
Saya	1
sedang	1
belajar	1
hitung	2
tf	1
idf	1

Menghitung IDF :

$$IDF_t = \log\left(\frac{N}{df_t}\right)$$

Term (t)	DF	IDF
Saya	1	log(2/1)= 0,301
sedang	1	log(2/1)= 0,301
belajar	1	log(2/1)= 0,301
hitung	2	log(2/2)=0
tf	1	log(2/1)= 0,301
idf	1	log(2/1)= 0,301

Menghitung TF-IDF :

Term (t)	D1	IDF	TF-IDF
			D1
Saya	1	log(2/1)= 0,301	0,301
sedang	1	log(2/1)= 0,301	0,301
belajar	1	log(2/1)= 0,301	0,301
hitung	2	log(2/2)=0	0

tf	1	log(2/1)= 0,301	0,301
idf	1	log(2/1)= 0,301	0,301

6.2.7. Pembobotan *Lexicon Based Features*

Lexicon Based Features merupakan fitur kata yang terdapat sentiment positif dan negatif bedasarkan kamus (*lexicon*). Lexicon merupakan kumpulan kata pada sentimen yang telah diketahui dan dihimpun dalam bentuk dataset [9].

Untuk melakukan proses pembobotan menggunakan fitur ini, dibutuhkan kamus (*lexicon*) yang mengandung kata yang sudah diberi sentimen.

6.2.8. *Support Vector Machine* (SVM)

Support Vector Machine adalah metode klasifikasi yang menggunakan cara mengklasifikasikan secara linear dengan menemukan *hyperlane* yang terbaik yang berfungsi sebagai pemisah antara 2 kelas. Prinsip dasarnya dilakukan pengklasifikasian secara liniar lalu dikembangkan sampai dapat dipakai pada permasalahan non linier dengan memasukkan konsep kernel trick pada ruang kerja berdimensi tinggi [10].

Alur kerja Support Vector Machine sebagai berikut :

1. Memetakan data
2. Meminimalisir nilai margin

   Dengan rumus :

   $$\frac{1}{2}\|w\|^2 - \frac{1}{2}(w_1^2 + w_2^2)$$

   Dengan syarat :

   $$y_i(X_1.w + b) - 1 \geq 0, i = 1,2,3,4,\dots,n$$

   $$y_i(X_1.W_1 + X_2.W_2 + b) \geq 1$$

3. Mencari persamaan *hyperlane*

4. Memetakan *hyperlane*
5. Melakukan pengujian terhadap data
6. Melakukan klasifikasi

Contoh :

Terdapat data seperti berikut (4 titik dari 2 kelas yang berbeda) :

X_1	X_2	Kelas(y)
1	1	1
1	-1	-1
-1	1	-1
-1	-1	-1

Hitunglah persamaan *hyperplane* data tersebut

Jawaban :

$y_i(X_1.w + b) - 1 \geq 0, i = 1,2,3,4, \dots, n$

$y_i(X_1.W_1 + X_2.W_2 + b) \geq 1$

Sehingga : $(W_1 + W_2 + b) \geq 1$ untuk $y_1 = 1$, $X_1=1$, $X_2=1$

$(-W_1 + W_2 - b) \geq 1$ untuk $y_2 = -1$, $X_1=1$, $X_2=-1$

$(W_1 - W_2 - b) \geq 1$ untuk $y_3 = -1$, $X_1=-1$, X2=1

$(W_1 + W_2 - b) \geq 1$ untuk $y_1 = -1$, $X_1=-1$, X2=-1

Lalu setelah itu dilakukan menjumlahkan / mengurangi masing persamaan yaitu persamaan 1 dan 2, 2 dan 3, serta 1 dan 3 sehingga menghasilkan nilai sebagai berikut :

$W_1 = 1$

$W_2 = 1$

$b = -1$

Sehingga dapat dicari persamaan dari *hyperplane*-nya

$W_1.X_1+W_2.X_2+b=0$

$1.X_1+1.X_2+-1=0$

$X_1+X_2-1=0$

$X_2 = 1-X_1$

6.2.9. *K-Nearest Neighbour* (KNN)

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*) [11].

Alur dari K-Nearest Neighbour sebagai berikut :

1. Menghitung jarak kesemua data *training* menggunakan *cosine similarity* atau *euclidean distance.*
2. Mengurutkan berdasarkan jarak terdekat dan ambil sejumlah K.
3. Mengambil K yang terbaik.
4. Mengambil label K terbaik sebelumnya yang paling banyak.

Contoh soal menggunakan *euclidean distance* :

Diberikan data sebagai berikut :

Tinggi	Berat	Jenis Kelamin
155	50	Perempuan
175	63	Laki - Laki
160	55	Perempuan
177	68	Laki - Laki
163	52	Perempuan
176	78	Laki - Laki

Tentukan jenis kelamin jika tinggi 172 dan berat 58 dengan K=3!

Jawaban :

Data 1 = $\sqrt{(155-172)^2+(50-58)^2} = 18{,}78829423$

Data 2 = $\sqrt{(175-172)^2+(63-58)^2} = 5{,}830951895$

Data 3 = $\sqrt{(165-172)^2+(55-58)^2} = 12{,}36931688$

Data 4 = $\sqrt{(177-172)^2+(68-58)^2} = 11{,}18033989$

Data 5 = $\sqrt{(163-172)^2+(52-58)^2} = 10{,}81665383$

Data 6 = $\sqrt{(176-172)^2+(78-58)^2} = 20{,}39607805$

Jika K=3 maka data yang diambil :

1. Data 6 (Laki - Laki)

2. Data 1 (Perempuan)

3. Data 3 (Laki -Laki)

Dan dapat disimpulkan jika K=3 maka prediksinya adalah Laki-Laki.

Persamaan dari cosine similarity ditunjukkan pada gambar dibawah ini.

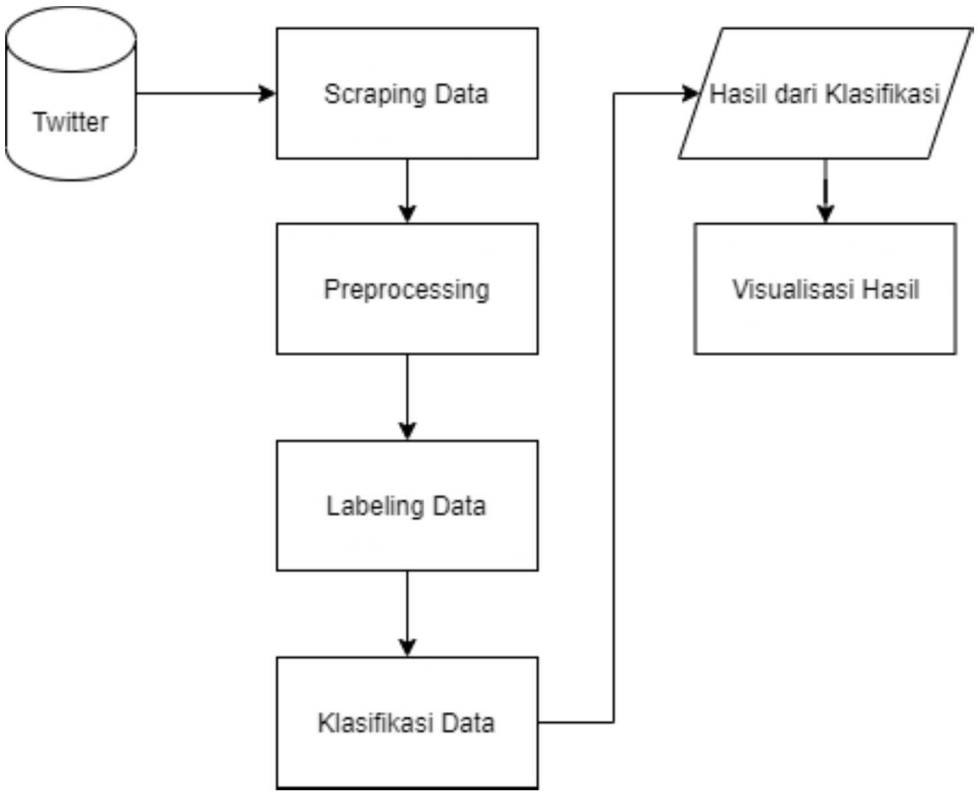
$$CosSim(q, d_j) = \frac{d_j . q}{|d_j| . |q|} = \frac{\sum_{i=1}^{t}(w_{ij} . w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \cdot \sqrt{\sum_{i=1}^{t} w_{iq}^2}}$$

Keterangan :

$CosSim(q,d_j)$	: Nilai kemiripan antara dokumen uji (q) dengan dokumen latih ke j (j_d)
t	: Jumlah term (kata)
d	: Dokumen
q	: Kata kunci (*query*)
W_{ij}	: Bobot term (kata) ke i pada dok. latih j
W_{iq}	: Bobot term (kata) ke i pada dok.uji q

7. Metode Penelitian

Untuk Menyusun tugas akhir, penulis menggunakan *flowchart* sebagai berikut :



Berikut langkah-langkah penyelesaian penelitian ini yaitu:

7.1. Pengumpulan Data

Pengumpulan data menggunakan teknik Sraping data dari Twitter. Data yang dikumpulkan berupa tweet dengan kata kunci “Harga Telur Turun” dan “Harga Telor Turun” dalam rentang waktu 29 Juli 2021 hingga 15 Oktober 2021 dan tidak disertakan posting *retweet.*

7.2. Pengolahan Data

Setelah melakukan pengumpulan data sebelum dianalisis perlu dilakukan proses awal atau dikenal dengan istilah Preprocessing. Proses ini akan mengolah data awal yang masih tidak beraturan untuk dijadikan data teratur yang dapat diterapkan pada proses selanjutnya. Preprocessing yang dilakukan terdiri dari Case Folding, Cleansing, Tokenizing, Normalization, dan Stopword Removing.

7.2.1. Case Folding

Case Folding adalah langkah untuk melakukan perubahan huruf besar atau huruf kapital (*uppercase*) yang terdapat pada teks menjadi huruf kecil (*lowercase*).

7.2.2. Cleansing

Cleansing adalah langkah membersihkan data dari hal – hal yang tidak perlu seperti URL, *hashtag*, tanda baca, angka dan lain sebagainya.

7.2.3. Tokenizing

Tokenizing adalah melakukan perubahan dari suatu kata pada kalimat yang dipisahkan oleh separator (*space*) menjadi sebuah token.

7.2.4. Normalization

Normalization adalah suatu proses dimana kata yang tidak baku atau singkat dirubah menjadi kata baku yang benar.

7.2.5. Stopword Removing

Stopword Removing adalah proses dimana kata penghubung seperti yang, di, ke, dari yang tidak diperlukan pada proses analisis dibuang.

7.2.6. Stemming

Steming adalah suatu kegiatan merubah kata yang memiliki imbuhan menjadi kata dasar. Perubahan kata dilakukan dengan menghilangkan prefix dan suffix. Stemming pada penelitian ini menggunakan *library* Sastrawi.

7.3. Labeling Data

Setelah data dibersihkan lalu dilakukan pelabelan pada data. Labeling pada data dilakukan secara otomatis menggunakan kamus yang sudah berisi bobot sentimen (*lexicon*) dan dihitung total dari sentimen bedasarkan jumlah bobot dari seluruh kata pada setiap data.

7.4. Pembobotan kata (TF-IDF)

Setelah di berikan label selanjutnya dilakukan pembobotan kata. Pembobotan kata dilakukan dengan menggunakan *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF).

7.5. Mengklasifikasikan Data

Proses ini bertujuan untuk mengolah data menjadi opini positif dan opini negatif. Ada banyak metode untuk mengklasifikasikan data, salah satunya adalah *Support Vector Machine* dan *K-Nearest Neighbour* . Merupakan salah satu metode untuk mengklasifikasikan data dan regresi. Pada penelitian ini, penulis menggunakan metode *Support Vector Machine* dan *K-Nearest Neighbour* untuk mengklasifikasikan data.

7.6. Visualisasi

Pada proses ini akan dilakukan visualisasi terhadap data yang dihasilkan dari proses klasifikasi. Tujuan dari proses ini untuk mempermudah membaca maksud dan informasi dari hasil analisis.

Daftar Referensi

[1] K. Makice, *Twitter API: Up and Running*. 2009.

[2] N. P. Aprilia, D. Pratiwi, and A. Barlianto, "Sentiment Visualization Of Covid-19 Vaccine Based On Naïve Bayes Analysis," vol. 6, no. 2, pp. 195–208, 2021.

[3] G. A. Buntoro, "Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter," *INTEGER J. Inf. Technol.*, vol. 1, no. 1, pp. 32–41, 2017, [Online]. Available: https://www.researchgate.net/profile/Ghulam_Buntoro/publication/316617194_Analisis_Sentimen_Calon_Gubernur_DKI_Jakarta_2017_Di_Twitter/links/5907eee44585152d2e9ff992/Analisis-Sentimen-Calon-Gubernur-DKI-Jakarta-2017-Di-Twitter.pdf.

[4] M. Lutz, *Python pocket ref*. 2014.

[5] A. Novantirani, M. K. Sabariah, and V. Effendy, "Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine," *e-Proceeeding Eng.*, vol. 2, no. 1, pp. 1–7, 2015.

[6] B. Zhao, "Encyclopedia of Big Data," *Encycl. Big Data*, no. May 2017, 2020, doi: 10.1007/978-3-319-32001-4.

[7] JustAnotherArchivist, "snscrape: A social networking service scraper in Python," 2021. https://github.com/JustAnotherArchivist/snscrape (accessed Oct. 22, 2021).

[8] J. Patterson and A. Gibson, *Deep learning: A Practionar Approach*, vol. 521, no. 7553. 2017.

[9] M. Desai and M. A. Mehta, "Techniques for sentiment analysis of Twitter data: A comprehensive survey," *Proceeding - IEEE Int. Conf. Comput. Commun. Autom. ICCCA 2016*, pp. 149–154, 2017, doi: 10.1109/CCAA.2016.7813707.

[10] R. Feldman and J. Sanger, *The Text Mining Handbook*. 2007.

[11] F. Gorunescu, *Data Mining Concepts,Models and Techniques*. 2011.