LINIER
LITERATUR INFORMATIKA & KOMPUTER
e-ISSN: 3063-2218
DOI : 10.33096/linier.v2i2.3104
Literatur Informatika & Komputer
Vol 2, No 2 (2025), pp 136-146

Implementasi Fitur *Vector Bag Of Word Dan TF IDF* untuk Analisis *Sentiment*

Muhammad Salman Al Markas[a], Siska Anraeni[b], Lutfi Budiman Ilmuwan[c]
Universitas Muslim Indonesia, Makassar, Indonesia
Email: [a]13020200267@umi.ac.id; [b]siska.anraeni@umi.ac.id; [c]lutfibudi.ilmawan@umi.ac.id

Received: 21-08-2024 | Revised: 11-05-2025 | Accepted: 11-06-2025 | Published: 29-06-2025

Abstrak

Penggunaan internet dengan media sosial mempengaruhi masyarakat terhadap kegiatan yang dilakukan saat ini. Salah satu media sosial yang sekarang ini sedang populer digunakan oleh masyarakat adalah X. Informasi yang disebarkan dapat merupakan berita, opini, komentar, serta kritikan. Data yang didapat dari tweet ini dapat menjelaskan tanggapan masyarakat terhadap pelayanan pajak dari X. Maka dari itu penelitian ini sangat efisien jika X menjadi media untuk pengambilan data mengenai komentar Masyarakat sehingga dapat memberikan efektivitas perubahan yang diberikan kepada instansi pemerintah. Analisis sentimen menjadi proses yang sangat penting dalam memahami isi data dengan tujuan mengolah komentar yang diberikan oleh pengguna melalui tweet di X mengenai sebuah produk, layanan, dan instansi. Karya ilmiah ini bertujuan untuk membandingkan fitur Vector Bag Of Word dan TF IDF untuk mengevaluasi seberapa penting suatu *term* dalam sebuah dokumen pada dokumen yang lebih besar. Seperti diketahui bahwa komputer hanya mampu memproses input yang numerik sehingga data opini public berupa teks perlu direpresentasikan sebagai nilai numerik yang dikenal dengan ekstraksi fitur dan dapat dilakukan menggunakan model Binary *Bag of Words* (BOW), Count BOW dan Term Frequency-Inverse Document Frequency (TF-IDF) dikarenakan kedua teknik tersebut sangat berperan baik dan sama-sama digunakan untuk merepresentasikan numerik dari data teks serta memiliki kekurangan dan kelebihan masing masing. Berdasarkan hasil analisis maka dapat disimpulkan dengan menganalisis statement dengan menggunakan Bag Of Word dan TF-IDF dapat mengetahui jumlah tiap kemunculan kata di setiap kalimat dan dari hasil yang didapatkan bahwa kata yang sering diucapkan dalam sentimen yaitu dengan bobot nilai TF-IDF sebesar 0.1403.

Kata kunci: *Bag Of Word,* Informasi, Komentar, *Sentiment, Term Frequency-Inverse Document Frequency*

Pendahuluan

Seiring dengan perkembangan teknologi saat ini, ilmu pengetahuan berbasis teknologi modern telah memenuhi kebutuhan dan mempermudah segala pekerjaan manusia [1]. Dengan adanya transformasi tersebut salah satu pengaruh paling besar yang bisa dirasakan adalah dalam bidang big data. Banyaknya pengguna media sosial membuat sarana komunikasi dalam bertukar informasi dijadikan sebagai tempat dalam berpendapat hingga bercerita dalam meluapkan yang dirasakan [2]. Media sosial telah mengubah cara kita berinteraksi dan berkomunikasi. Dalam beberapa tahun terakhir, media sosial telah menjadi alat penting dalam berpolitik dan aktivisme [3]. Menjadi wadah untuk individu serta kelompok menyebarkan pesan menyampaikan berbagai informasi, memobilisasi pendukung, dan berinteraksi langsung dengan masyarakat, mengorganisir protes dan kampanye, serta mempengaruhi opini public [4]. Media sosial bisa menjadi alat untuk mengetahui atau menemukan sentimen publik terhadap suatu sosok. Pemerintah dalam hal ini dapat menjadi objek yang dianalisis untuk melihat sentimen masyarakat terhadap pemerintahan baik dari segi kinerja maupun kebijakan yang diambil [5].

Penggunaan internet dengan media sosial mempengaruhi masyarakat terhadap kegiatan yang dilakukan saat ini. Salah satu media sosial yang sekarang ini sedang populer digunakan oleh masyarakat adalah X [6]. Pada laporan finansial X quarter 2019 ke-4, pengguna aktif harian di platform X dicatat ada 145 juta pengguna [7], dan di Indonesia sendiri menjadi salah satu negara dengan penggunaan terbanyak dalam pengguna aktif X [8]. Informasi yang disebarkan dapat merupakan berita, opini, komentar, kritikan, dan netral. Data yang didapat dari tweet ini dapat menjelaskan tanggapan masyarakat terhadap pelayanan pajak dari X. Maka dari itu penelitian ini sangat efisien jika X menjadi media untuk pengambilan data mengenai komentar masyarakat terhadap instansi pemerintah sehingga dapat memberikan efektivitas perubahan yang diberikan kepada instansi pemerintah. Informasi yang dikumpulkan dari data tweet merupakan data teks yang tidak terstruktur.

Analisis sentimen merupakan metode klasifikasi yang digunakan untuk mengelompokkan opini yang terkandung dalam sebuah teks [9]. Analisis sentimen sering digunakan untuk mengetahui opini masyarakat

melalui ulasan atau komentar mengenai instansi pemerintah [10]. Untuk mengenali dan mengekstraksi opini dalam bentuk teks maka dilakukan analisis sentimen guna mengubah informasi yang tadinya tidak terstruktur dapat diubah menjadi data yang lebih terstruktur. Analisis sentimen merupakan salah satu cabang dari text mining yang melakukan identifikasi teks dan kemudian mengekstrak informasi dari teks yang telah diidentifikasi menjadi informasi subjektif dalam sumber [11]. Analisis sentimen menjadi proses yang sangat penting dalam memahami isi data dengan tujuan mengolah komentar yang diberikan oleh si pengguna melalui tweet di X mengenai sebuah produk, layanan, dan sebuah instansi [12].

Metode analisis sentimen dalam Bahasa Indonesia sudah banyak dilakukan menggunakan metode classical machine learning seperti Naïve Bayes, Maximum Entropy (ME), Support Vector Machine (SVM), dan Decision Tree [13]. Beberapa penelitian belum menggunakan perbandingan metode ekstraksi fitur dalam penilaian kinerjanya. Kinerja klasifikasi dapat dipengaruhi oleh vektor fitur atau ekstraksi fitur ke dalam algoritma klasifikasi [14]. Seperti diketahui bahwa komputer hanya memproses input numerik sehingga data opini public berupa teks perlu direpresentasikan sebagai nilai numerik yang dikenal dengan ekstraksi fitur dan dapat dilakukan menggunakan model *Binary Bag of Words* (BOW), Count (BOW) dan (Term Frequency-Inverse Document Frequency) TF-IDF dikarenakan kedua ekstraksi fitur tersebut sangat berperan baik dan digunakan untuk merepresentasikan numerik dari data teks serta memiliki kekurangan dan kelebihan masing masing [15][16]. Karya ilmiah ini bertujuan untuk membandingkan fitur vector Bag Of Word dan TF IDF untuk mengevaluasi seberapa penting suatu kata (term) dalam sebuah dokumen dalam konteks koleksi dokumen yanglebih besar. Karya ilmiah ini akan menganalisis sentimen yang mengkaji tentang komentar masyarakat terhapat instansi pemerintah yang dikembangkan dalam bentuk karya ilmiah yang Berjudul Fitur Vector Bag Of Word dan TF IDF Untuk Analisis Sentiment di akun x instansi pemerintah. Diharapkan karya ilmiah ini dapat diolah sehingga memiliki representasi yang akurat dari setiap dokumen serta mengorganisir informasi tersebut dalam ke dalam proses klasifikasi.

Metode

A. Pengumpulan Data
Pengumpulan data yang dilakukan dalam analisis sentimen adalah pengumpulan data yang diambil dari kolom komentas X di akun instansi pemerintah. Data ini diambil dimulai per-Maret 2024.

B. Labeling Data
Pada tahap ini adalah tahap yang memberikan label kategori sentimen pada data yang akan dianalisis. Misalnya, data dari X diamati dan diberikan sentimen positif, negatif, atau netral. Dalam pengerjaan tahap ini ada beberapa daftar rules dan polaritasnya untuk mempermudah dalam pemberian label. Contohnya seperti “jika ada data yang sekilas menyatakan ajakan kebersamaan atau dukungan, tetapi disertai dengan ekspektasi yang berlawanan dan di mana si penulis pun terbukti secara logika tidak akan melakukan ajakan tersebut, kita akan padankan data tersebut dengan polaritas negatif dengan konteks menyindir”. Selain itu, ada juga panduan konteks kalimat tanya, adjectives, how to differ context, dan lain-lain. Pemberian label dalam penelitian ini juga dilakukan secara manual dan membutuhkan waktu satu minggu untuk memberi label pada 500 data tweet.

Tabel 1. Hasil *Labeling* Manual

Username	Komentar	Keterangan
@aviantiarmand	Aku tidak rela uang pajak yang aku bayarkan dipakai untuk menggaji anggota DPR yang seperti ini. RT yang setuju.	Negatif
@podoradong	Kami mengajak sobat semua untuk berdoa Mari kita doakan untuk para anggota dewan yang membohongi mengakali hingga culas pada rakyat dalam pembahasan RUU ciptaker dengan alasan listrik DPR padam semoga dilaknat dunia akhirat Al Fatihah	Negatif
@dr_koko28	Para wanita harus tahu ini. Semakin banyak mereka mengkonsumsi minuman manis semakin besar risiko mereka mengalami PCOS. Banyak yang infertil gara-gara itu. Titipkanlah ke kaum wanita di	Positif

	DPR untuk perjuangkan agar batas jumlah gula dalam produk lebih diturunkan.	
@muluojanmulu	Jgn ngeluh jd beban keluarga dpr aja santai jd beban negara.	Negatif

Pada Tabel 1 merupakan hasil labeling manual, dapat diketahui bahwa sebagian besar cuitan yang dianalisis memiliki sentimen negatif terhadap kinerja atau perilaku anggota DPR.

C. Pra-proses Data

Setelah data diberi label maka langkah selanjutnya dengan melakukan preprocessing pada data preprocessing. Pre-Processing adalah proses awal dilakukannya perbaikan untuk menghilangkan noise yang melibatkan empat tahap, mulai dari data cleaning dan case folding, yang membersihkan data dari kesalahan termasuk noise data dan mengubah semua teks menjadi huruf kecil [17]. Kedua, filtering, yang menghapus kata-kata atau karakter yang tidak relevan juga menghilangkan slang dan kata-kata pendek. Ketiga, stemming, yang mengambil kata dasar kata untuk membantu dalam klasifikasi teks. Keempat, tokenizing, yang mengubah kalimat menjadi makna yang lebih spesifik menggunakan konsep bigram, yaitu dua kata yang berpasangan [18].

D. Ekstraksi Fitur

Selanjutnya, yang perlu dilakukan adalah mengubah teks menjadi representasi numerik yang dapat diproses oleh algoritma pembelajaran mesin. Beberapa teknik ekstraksi fitur yang umum digunakan dalam analisis sentimen adalah:

1. Bag of Words (BoW): Menghitung frekuensi kemunculan kata-kata dalam setiap teks [19].
2. Term Frequency-Inverse Document Frequency (TF-IDF) : Menghitung bobot kata berdasarkan frekuensi kemunculannya dalam teks dan keseluruhan dokumen [20].

Rumus metode Term Frequency-Inverse Document Frequency (TF-IDF) [21]:

$$tf = 0.5 + 0.5\, x\, d(x,y) = \frac{tf}{max\ (tf)} \quad (1)$$

$$idf_t = \log \frac{D}{df_t} \quad (2)$$

$$W_{d,t} = tf_{d,t}\, x\, idf_{d,t} \quad (3)$$

Keterangan:

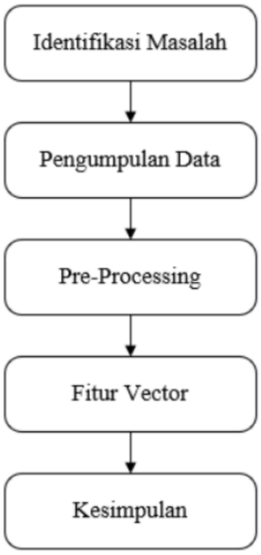
tf = banyaknya kata yang dicari dalam dokumen
$max\ (tf)$ = jumlah kemunculan terbanyak kata di dalam dokumen yang sama
d = total dokumen
df_t = jumlah dokumen yang mengandung term t
idf_t = *Inversed Document Frequency*
W = bobot dokumen ke-d terhadap kata t

Perancangan

Tahapan penelitian ini berisikan tahapan-tahapan penelitian yang akan dilakukan. Adapun tahapan penelitian dapat di jelaskan pada Gambar 1:

A. Identifikasi Masalah, tahap pertama yang dilakukan untuk mengidentifikasi masalah yang terjadi pada komentar di akun X instansi pemerintah.
B. Pengumpulan Data, Pengumpulan data dilakukan menggunakan aplikasi X untuk memperoleh data yang diperlukan selama penelitian.
C. Pre-Processing, pada tahap preprocessing ini, data yang telah di crawling melalui Youtube API akan diolah menjadi data yang dapat digunakan pada tahapan selanjutnya. Tahapan Preprocessing terdiri dari case folding, tokenizing, filtering dan stemming.

D. Fitur vector, pada tahap ini dilakukan perhitungan frekuensi kemunculan kata-kata dalam setiap teks dan pembobotan kata berdasarkan jumlah kemunculan kata.
E. Kesimpulan, kemudian hasil akhir diberikan kesimpulan dengan menjawab rumusan masalah yang telah dijabarkan sebelumnya.



Gambar 1. Tahapan Penelitian

Pemodelan

Data yang digunakan pada penelitian ini merupakan hasil crawling yang dilakukan di X Application Programming Interface (API) sebanyak 500 komentar pengguna X. Dataset akan diambil secara imbang berdasarkan pembagian kelas positif, negatif dan netral. Opini positif yaitu mengidentifikasikan bahwa seseorang memiliki perilaku yang baik atau menyukai suatu postingan. Ini mencerminkan sifat positif atau penilaian menguntungkan terhadap subjek tertentu. Sedangkan opini negatif menunjukkan perilaku yang kurang baik atau tidak menyukai suatu postingan. Ini mencerminkan sikap negatif penelitian yang merugikan atau menyudutkan subjek tertentu. Adapun opini netral yaitu memperlihatkan perilaku yang seimbang antara berbagi perspektif atau sudut pandang yang ada. Pada penelitian ini, data yang digunakan adalah data primer dengan menggunakan dataset hasil crawling menggunakan Youtube Application Programming Interface (API). Data tersebut akan diolah menggunakan Pyhton Yang dimana tahapan preprocessing dalam hal ini meliputi :

A. *Pre-Processing*

```
# crawl data

filename = 'DPR lang:id'
limit = 500

!npx  --yes tweet-harvest@2.6.1 -o "{filename}" -s "{search_keyword}" -l {limit} --token {twitter_auth_token}
```

Gambar 2. Proses *Crawling*

Gambar 2 merupakan proses crawling di google collab didapatkan sebanyak 500 data yang berhasil di crawling dari X yang berhubungan dengan komentar di instansi pemerintah, yang dimana hasil tersebut didapatkan lebih banyak teks komentar negative dibandingkan komentar positif.

1. Proses *Case Folding*
   Mengubah semua huruf dalam dokumen menjadi huruf kecil.

```
def case_folding(data):

    data = data.lower()
    data = data.replace('\n',' ')
    data = re.sub('[^\w\s]+', ' ',data)
    data = re.sub('\d+', '', data)
    return data

df["full_text"] = df["full_text"].apply(case_folding)
df
```

	full_text
0	dpr live https t co ljmuadnub
1	dpr ian https t co xljhvklf
2	dpr salah server region https t co ltbauyebh
3	goblok e anggota dpr ojo sampek ganggu jiwa bu...
4	gedung dpr diliat liat kaya memek https t co m...
...	...
499	paling gk suka klo lg nongkrong tbtb dituduh a...
500	ilckpkdiperkuatataudiperlemah narasumber il...
501	srikandi nasdem yang masuk besar perempuan p...
502	raker dengan komisi ii dpr waseskab berharap u...
503	anggota mpr yang dilantik terdiri dari anggot...

504 rows × 1 columns

Gambar 3. Proses *Crawling*

Pada Gambar 3 merupakan proses *case folding* dan outpunya mengubah semua huruf di dokumen menjadi huruf kecil.

Tabel 2. Penerapan *Case Folding*

Sebelum	Sesudah
Anggota MPR yang dilantik terdiri dari 575 anggota DPR dan 136 anggota DPD. Pelantikan diawali dengan pembacaan Keputusan Presiden oleh Sekretaris Jenderal MPR RI dan dilanjutkan pengucapan sumpah janji jabatan yang dipandu oleh Ketua MA Muhammad Hatta Ali. #KetuaMPR #MPR https://t.co/b5e3WFgueb	Anggota mpr yang dilantik terdiri dari 575 anggota dpr dan 136 anggota dpd. pelantikan diawali dengan pembacaan keputusan presiden oleh sekretaris jenderal mpr ri dan dilanjutkan pengucapan sumpah janji jabatan yang dipandu oleh ketua ma muhammad hatta ali. #ketuampr #mpr https://t.co/b5e3wfgueb

Tabel 2 merupakan penerapan *case folding* yang menampilkan kata sebelum dan sesudah *case folding*.

2. Proses *Case Folding*

Pemisahan kata dalam suatu kalimat menjadi token-token/bagian-bagian tertentu.

```
from nltk.tokenize import word_tokenize
def tokenisasi(text):
  return word_tokenize(text)


df["fulltext_token"] = df["full_text"].apply(tokenisasi)
df
```

Gambar 4. Proses *Crawling*

Gambar 4 merupakan proses *tokenizing* dan outputnya yang melakukan pemisahan kata dalam suatu kalimat menjadi token-token/bagian-bagian tertentu.

Tabel 3. Penerapan *Tokenizing*

Sebelum	Sesudah
DPR salah server region https://t.co/3LtBAUyeBH	DPR, salah, server, region, https,t,co,ltbauyebh

Tabel 3 merupakan penerapan *tokenizing* yang menampilkan kata sebelum dan sesudah *tokenizing*.

3. Proses *Normalization*

Proses mengubah kata tidak standar atau ditingkatan menjadi kata formal yang bisa dipakai sehari-hari. Adapun normalization diambil dari github dari judul text-processing [22].

```
import pandas as pd

kamus= pd.read_csv("/content/drive/MyDrive/DataSet/slang.csv")
kamus = kamus.filter(['slang', 'formal'], axis=1)

dict_normalisasi = dict(kamus.values)

def normalisasi(document):

    return [
        dict_normalisasi[token] if token in dict_normalisasi else token
        for token in document
    ]

df["Review_normalisasi"] = df["Reviews_token"].apply(normalisasi)
df
```

	full_text	fulltext_token	Review_normalisasi
0	DPR LIVE https://t.co/L3JMuAdnuB	[DPR, LIVE, https, :, //t.co/L3JMuAdnuB]	[DPR, LIVE, https, :, //t.co/L3JMuAdnuB]
1	DPR IAN https://t.co/1XlJHv5klF	[DPR, IAN, https, :, //t.co/1XlJHv5klF]	[DPR, IAN, https, :, //t.co/1XlJHv5klF]
2	DPR salah server region https://t.co/3LtBAUyeBH	[DPR, salah, server, region, https, :, //t.co/...	[DPR, salah, server, region, https, :, //t.co/...
3	Goblok e anggota DPR ojo sampek ganggu jiwa bu...	[Goblok, e, anggota, DPR, ojo, sampek, ganggu,...	[Goblok, e, anggota, DPR, jangan, sampai, gang...
4	Gedung DPR diliat liat kaya memek https://t.co...	[Gedung, DPR, diliat, liat, kaya, memek, https...	[Gedung, DPR, dilihat, liat, kaya, memek, http...
...	...	...	...
499	paling gk suka klo lg nongkrong tbtb dituduh a...	[paling, gk, suka, klo, lg, nongkrong, tbtb, d...	[paling, tidak, suka, kalau, lagi, nongkrong, ...
500	#ILCKPKDiperkuatAtauDiperlemah | NARASUMBER IL...	[#, ILCKPKDiperkuatAtauDiperlemah, |, NARASUMB...	[#, ILCKPKDiperkuatAtauDiperlemah, |, NARASUMB...
501	3 Srikandi NasDem yang masuk 10 besar perempua...	[3, Srikandi, NasDem, yang, masuk, 10, besar, ...	[3, Srikandi, NasDem, yang, masuk, 10, besar, ...
502	Raker Dengan Komisi II DPR Waseskab Berharap U...	[Raker, Dengan, Komisi, II, DPR, Waseskab, Ber...	[Raker, Dengan, Komisi, II, DPR, Waseskab, Ber...
503	Anggota MPR yang dilantik terdiri dari 575 ang...	[Anggota, MPR, yang, dilantik, terdiri, dari, ...	[Anggota, MPR, yang, dilantik, terdiri, dari, ...

504 rows × 3 columns

Gambar 5. Proses *Normalization*

Gambar 5 merupakan proses normalization dan outputnya yang mengubah kata tidak standar atau ditingkatan menjadi kata formal.

Tabel 4. Penerapan *Normalization*

Sebelum	Sesudah
Paling gk suka klo lg nongkrong tbtb dituduh anggota dpr	Paling,tidak,suka,kalau,lagi,nongkrong, tiba-tiba,dituduh,anggota,dpr

Tabel 4 merupakan penerapan *normalization* yang menampilkan kata sebelum dan sesudah *normalization*.

4. Proses *Stemming*

Stemming adalah proses mengubah kata-kata yang berimbuhan menjadi ke kalimat dasar.

```
[22] factory = StemmerFactory()
     stemmer = factory.create_stemmer()


     def stemming(term):
         return stemmer.stem(term)

     df["Reviews_clean"] = [' '.join(map(str, l)) for l in df["Review_normalisasi"]]
     df["Reviews_stem"] = df["Reviews_clean"].apply(stemming)
     df
```

	full_text	fulltext_token	Reviews_normalisasi	Reviews_stopword	Reviews_clean	Reviews_stem
0	dpr live https t co ijmuadnub	[dpr, live, https, t, co, ijmuadnub]	[dpr, live, https, t, co, ijmuadnub]	[dpr, live, https, t, co, ijmuadnub]	dpr live https t co ijmuadnub	dpr live https t co ijmuadnub
1	dpr ian https t co xljhvkif	[dpr, ian, https, t, co, xljhvkif]	[dpr, ian, https, t, co, xljhvkif]	[dpr, ian, https, t, co, xljhvkif]	dpr ian https t co xljhvkif	dpr ian https t co xljhvkif
2	dpr salah server region https t co ttbauyebh	[dpr, salah, server, region, https, t, co, ttb...	[dpr, salah, server, region, https, t, co, ttb...	[dpr, salah, server, region, https, t, co, ttb...	dpr salah server region https t co ttbauyebh	dpr salah server region https t co ttbauyebh
3	goblok e anggota dpr ojo sampek ganggu jiwa bu...	[goblok, e, anggota, dpr, ojo, sampek, ganggu...	[goblok, e, anggota, dpr, jangan, sampai, gang...	[goblok, e, anggota, dpr, ganggu, jiwa, bucinm...	goblok e anggota dpr ganggu jiwa bucinmu https...	goblok e anggota dpr ganggu jiwa bucinmu https...
4	gedung dpr dilat liat kaya memek https t co m...	[gedung, dpr, dilat, liat, kaya, memek, https...	[gedung, dpr, dilihat, lihat, kaya, memek, htt...	[gedung, dpr, lihat, kaya, https, t, co, mvulo...	gedung dpr lihat kaya https t co mvulovivd	gedung dpr lihat kaya https t co mvulovivd
...	...	...	...	...	...	...
499	paling gk suka klo lg nongkrong tbtb dituduh a...	[paling, gk, suka, klo, lg, nongkrong, tbtb, d...	[paling, tidak, suka, kalau, lagi, nongkrong, ...	[suka, nongkrong, dituduh, anggota, dpr]	suka nongkrong dituduh anggota dpr	suka nongkrong tuduh anggota dpr
500	ilcikpidperluatataudiperlemah narasumber ii...	[ilcikpidperluatataudiperlemah, narasumber, ii...	[ilcikpidperluatataudiperlemah, narasumber, ii...	[ilcikpidperluatataudiperlemah, narasumber, ii...	ilcikpidperluatataudiperlemah narasumber ilc a...	ilcikpidperluatataudiperlemah narasumber ilc a...
501	srikandi nasdem yang masuk besar perempuan p...	[srikandi, nasdem, yang, masuk, besar, perempu...	[srikandi, nasdem, yang, masuk, besar, perempu...	[srikandi, nasdem, masuk, perempuan, perah, s...	srikandi nasdem masuk perempuan perah suara d...	srikandi nasdem masuk perempuan raih suara dpr...
502	raker dengan komisi ii dpr waseskab berharap u...	[raker, dengan, komisi, ii, dpr, waseskab, ber...	[raker, dengan, komisi, ii, dpr, waseskab, ber...	[raker, komisi, ii, dpr, waseskab, berharap, u...	raker komisi ii dpr waseskab berharap usulan p...	raker komisi ii dpr waseskab harap usul tambah...
503	anggota mpr yang dilantik terdiri dari anggot...	[anggota, mpr, yang, dilantik, terdiri, dari, ...	[anggota, mpr, yang, dilantik, terdiri, dari, ...	[anggota, mpr, dilantik, anggota, dpr, anggota...	anggota mpr dilantik anggota dpr anggota dpd p...	anggota mpr lantik anggota dpr anggota dpd lan...

504 rows × 6 columns

Gambar 6. Proses *Stemming*

Gambar 6 merupakan proses *stemming* dan outpunya proses mengubah kata-kata yang berimbuhan menjadi ke kalimat dasar menggunakan *library Sastrawi.Stemmer.StemmerFactory* yang di import dari *phyton*.

Tabel 5. Penerapan *Stemming*

Sebelum	Sesudah
ketua dpr puan maharani dilaporkan mahkamah kehormatan dewan mkd terkait dugaan pelanggaran etik perayaan ulang tahunnya sidang paripurna September laporan dilayangkan kaukus muda anti korupsi kamaks	ketua dpr puan maharani lapor mahkamah hormat dewan mkd kait duga langgar etik raya ulang tahun sidang paripurna september lapor layang kaukus muda anti korupsi kamaksi

Tabel 5 merupakan penerapan proses *stemming* dengan mengubah kata berimbuhan menjadi kata dasar yang menampilkan kata sebelum dilakukan *stemming* dan sesudah *stemming.*

5. Proses *Stopword Removal*

Stopword Removal adalah proses menghilangkan kata-kata yang tidak dibutuhkan atau kata-kata yang tidak memiliki arti. Adapun *stopword github* dari judul *text-processing* [22].

```
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemover, ArrayDictionary
import string
def stop_stem(text):
    #stopword
    with open('/content/drive/MyDrive/DataSet/kamus.txt') as kamus:
        word = kamus.readlines()
        list_stopword = [line.replace('\n',"") for line in word]
    dictionary = ArrayDictionary(list_stopword)
    stopword = StopWordRemover(dictionary)
    text = stopword.remove(text)

    return text

df["Reviews_stopword"] = df["Reviews_stem"].apply(stop_stem)
df
```

	full_text	Reviews_token	Review_normalisasi	Reviews_clean	Reviews_stem	Reviews_stopword
0	dpr live https t co ljmuadnub	[dpr, live, https, t, co, ljmuadnub]	[dpr, live, https, t, co, ljmuadnub]	dpr live https t co ljmuadnub	dpr live https t co ljmuadnub	dpr live https co ljmuadnub
1	dpr ian https t co xljhvklf	[dpr, ian, https, t, co, xljhvklf]	[dpr, ian, https, t, co, xljhvklf]	dpr ian https t co xljhvklf	dpr ian https t co xljhvklf	dpr ian https co xljhvklf
2	dpr salah server region https t co ltbauyebh	[dpr, salah, server, region, https, t, co, ltb...	[dpr, salah, server, region, https, t, co, ltb...	dpr salah server region https t co ltbauyebh	dpr salah server region https t co ltbauyebh	dpr salah server region https co ltbauyebh
3	goblok e anggota dpr ojo sampek ganggu jiwa bu...	[goblok, e, anggota, dpr, ojo, sampek, ganggu,...	[goblok, e, anggota, dpr, ojo, sampai, ganggu,...	goblok e anggota dpr ojo sampai ganggu jiwa bu...	goblok e anggota dpr ojo sampai ganggu jiwa bu...	goblok anggota dpr ojo sampai ganggu jiwa buci...
4	gedung dpr dillat liat kaya memek https t co m...	[gedung, dpr, dillat, liat, kaya, memek, https...	[gedung, dpr, dillat, lihat, seperti, memek, h...	gedung dpr dillat lihat seperti memek https t ...	gedung dpr liat lihat seperti memek https t co...	gedung dpr liat lihat seperti memek https co m...
...	...	...	...	...	...	...
499	paling gk suka klo lg nongkrong tbtb dituduh a...	[paling, gk, suka, klo, lg, nongkrong, tbtb, d...	[paling, tidak, suka, kalau, lagi, kumpul, tbt...	paling tidak suka kalau lagi kumpul tbtb ditud...	paling tidak suka kalau lagi kumpul tbtb tuduh...	paling tidak suka kalau kumpul tbtb tuduh angg...
500	ilckpkdiperkuatataudiperlemah narasumber il...	[ilckpkdiperkuatataudiperlemah, narasumber, il...	[ilckpkdiperkuatataudiperlemah, narasumber, il...	ilckpkdiperkuatataudiperlemah narasumber ilc a...	ilckpkdiperkuatataudiperlemah narasumber ilc a...	ilckpkdiperkuatataudiperlemah narasumber ilc a...
501	srikandi nasdem yang masuk besar perempuan p...	[srikandi, nasdem, yang, masuk, besar, perempu...	[srikandi, nasdem, yang, masuk, besar, perempu...	srikandi nasdem yang masuk besar perempuan per...	srikandi nasdem yang masuk besar perempuan rai...	srikandi nasdem masuk besar perempuan raih sua...
502	raker dengan komisi ii dpr waseskab berharap u...	[raker, dengan, komisi, ii, dpr, waseskab, ber...	[raker, dengan, komisi, ii, dpr, waseskab, ber...	raker dengan komisi ii dpr waseskab berharap u...	raker dengan komisi ii dpr waseskab harap usul...	raker dengan komisi dpr waseskab harap usul ta...
503	anggota mpr yang dilantik terdiri dari anggot...	[anggota, mpr, yang, dilantik, terdiri, dari, ...	[anggota, mpr, yang, dilantik, terdiri, dari, ...	anggota mpr yang dilantik terdiri dari anggota...	anggota mpr yang lantik diri dari anggota dpr ...	anggota mpr lantik diri anggota dpr anggota dp...

504 rows × 6 columns

Gambar 7. Proses *Stopword*

Gambar 7 merupakan proses *stopword removal* dan outputnya yang menghilangkan kata-kata yang tidak dibutuhkan atau kata-kata yang tidak memiliki arti.

Tabel 6. Penerapan *Stopword*

Sebelum	Sesudah
raker dengan komisi II DPR waseskab berharap	raker komisi II DPR waseskab berharap

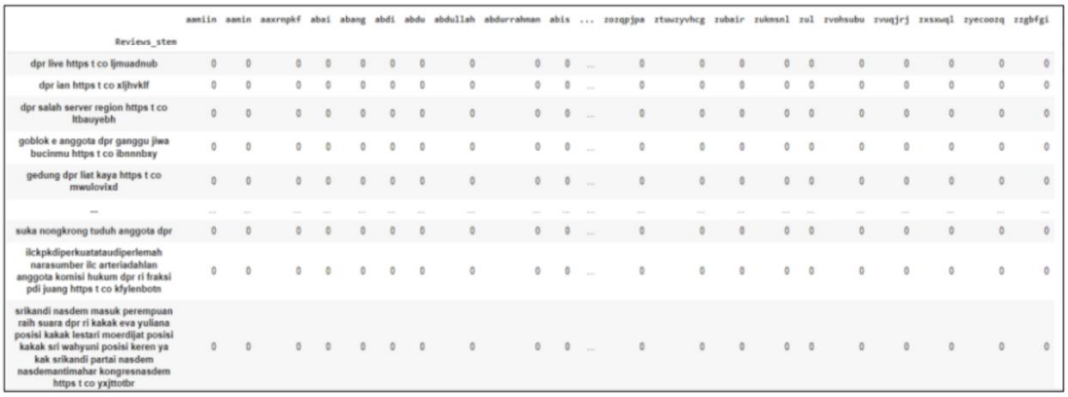
Tabel 6 merupakan tabel sebelum dan sesudah menggunakan stop removal.

B. Ekstraksi Fitur

1. *Bag Of Words*

Menghitung frekuensi kemunculan kata-kata dalam setiap teks.

```
from sklearn.feature_extraction.text import CountVectorizer
documents = df['Reviews_stem']
count_vector = CountVectorizer()
count_vector.fit(documents)
count_vector.get_feature_names_out()
doc_array = count_vector.transform(documents).toarray()
frequency_matrix = pd.DataFrame(doc_array,index=documents,columns=count_vector.get_feature_names_out())
frequency_matrix
```



Reviews_sten	aamiin	aamin	aaxrnpkf	abai	abang	abdi	abdu	abdullah	abdurrahman	abis	...	zozqpjpa	ztuwzyvhcg	zubair	zukmsnl	zul	zvohsubu	zvuqjrj	zxsnuql	zyecoozq	zzgbfgi
dpr live https t co ljmuadnub	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
dpr ian https t co xljhvklf	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
dpr salah server region https t co ltbauyebh	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
goblok e anggota dpr ganggu jiwa bucinmu https t co ibnnnbxy	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
gedung dpr liat kaya https t co mwulovlxd	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
suka nongkrong tuduh anggota dpr	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
ilckpkdiperkuatataudiperlemah narasumber ilc arteriadahlan anggota komisi hukum dpr ri fraksi pdi juang https t co kfylenbotn	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
srikandi nasdem masuk perempuan raih suara dpr ri kakak eva yuliana posisi kakak lestari moerdijat posisi kakak sri wahyuni posisi keren ya kak srikandi partai nasdem nasdemantimahar kongresnasdem https t co yxjttotbr	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Gambar 8. Proses *Bag Of Word*

Gambar 8 merupakan proses *Bag Of Words (BOW)* yang menghitung jumlah frekuensi kemunculan kata-kata dalam setiap teks dokumen.

2. TF-IDF

Menghitung bobot kata berdasarkan frekuensi kemunculannya dalam teks dan keseluruhan dokumen.

```
from sklearn.feature_extraction.text import TfidfVectorizer

documents = df['Reviews_stopword']
vectorizer = TfidfVectorizer()
X_tfidf = vectorizer.fit_transform(documents)


tfidf = TfidfVectorizer()
response = tfidf.fit_transform(Z)


feature_names = tfidf.get_feature_names_out()
for col in response.nonzero()[1]:
    print (feature_names[col], ' - ', response[0, col])
```

```
story  -  0.0
ig  -  0.0
ian  -  0.0
co  -  0.14260032207967865
https  -  0.14260032207967865
dpr  -  0.14031426634037622
mfr6jywakq  -  0.0
kemal  -  0.0
golkar  -  0.0
partai  -  0.0
kata  -  0.0
calon  -  0.0
co  -  0.14260032207967865
https  -  0.14260032207967865
dpr  -  0.14031426634037622
iloojjdibq  -  0.0
teradu  -  0.0
mengadili  -  0.0
memutuskan  -  0.0
wacana  -  0.0
persetujuan  -  0.0
ringan  -  0.0
menjatuhi  -  0.0
mengklaim  -  0.0
uud  -  0.0
amendemen  -  0.0
soesatyo  -  0.0
tertulis  -  0.0
teguran  -  0.0
sanksi  -  0.0
1945  -  0.0
mpr  -  0.0
politik  -  0.0
menyatakan  -  0.0
```

Gambar 9. Proses TF-IDF

Gambar 9 merupakan proses TF-IDF beserta outputnya yang menghitung bobot kata berdasarkan frekuensi kemunculannya dalam teks. Berdasarkan hasil TF-IDF kata dpr memiliki bobot nilai TF-IDF tinggi sebesar 0.1403.

Kesimpulan

Berdasarkan hasil analisis maka dapat disimpulkan dengan menganalisis statement dengan menggunakan *Bag Of Word* dan TF-IDF dapat mengetahui jumlah tiap kemunculan kata di setiap kalimat dan dari hasil yang didapatkan bahwa kata yang sering diucapkan dalam sentimen yaitu kata dpr dengan bobot nilai TF-IDF sebesar 0.1403.

Daftar Pustaka

[1] N. H. F. Mulyani, "Analisis perkembangan ilmu pengetahuan dan teknologi (IPTEK) dalam pendidikan," 2021, *academia.edu*. doi: 10.31004/jpdk.v3i1.1432.

[2] K. T. Putra and M. A. Hariyadi, "Perbandingan Feature Extraction Tf-Idf Dan Bow Untuk Analisis Sentimen Berbasis Svm," *J. Cahaya MAndalika 1449*, 2023.

[3] I. R. Putri and E. Pratiwi, "Aktivisme digital dan pemanfaatan media baru sebagai pendekatan pemberdayaan masyarakat atas isu lingkungan," *Bricol. J. Magister Ilmu Komun.*, 2022, doi: 10.30813/bricolage.v8i2.3303.

[4] A. A. Rahmawati, Metha Binety Maharani, and M. N. F. Al Amin, "Peran Media Sosial Dalam Proses Pengambilan Keputusan Politik Melalui Pendekatan Problem Tree Analysis," *ARIMA J. Sos. Dan Hum.*, vol. 1, no. 4 SE-Articles, pp. 112–121, Apr. 2024, doi: 10.62017/arima.v1i4.1043.

[5] M. Jurnal, "Sumber Daya Sektor Pelayanan Publik Era Revolusi Industri 4.0: Profesional Dan Komunikatif Sebuah Tantangan," *J. MSDA (Manajemen Sumber Daya Apar.*, 2020, doi: 10.33701/jmsda.v8i2.1404.

[6] M. F. Alfajri, V. Adhiazni, and Q. Aini, "Pemanfaatan Social Media Analytics Pada Instagram Dalam Peningkatan," *Interak. J. Ilmu Komun.*, 2019, doi: https://dx.doi.org/10.14710/interaksi.8.1.34-42.

[7] S. Juanita, "Analisis sentimen persepsi masyarakat terhadap pemilu 2019 pada media sosial twitter menggunakan naive bayes," 2020, *academia.edu*. doi: 10.30865/mib.v4i3.2140.

[8] K. Kartini, J. Syahrina, N. Siregar, and N. Harahap, "Penelitian Tentang Instagram," 2022.

[9] L. B. Ilmawan and M. A. Mude, "Perbandingan metode klasifikasi Support Vector Machine dan Naïve Bayes untuk analisis sentimen pada ulasan tekstual di Google Play Store," 2020, *academia.edu*. doi: 10.33096/ilkom.v12i2.597.154-161.

[10] H. Darwis, A. N. P. Pagala, and S. Anraeni, "Analysis of Public Sentiment about Childfree in Indonesia using Support Vector Machine Methods," *2025 19th Int. Conf. Ubiquitous Inf. Manag. Commun. IMCOM 2025*, 2025, doi: 10.1109/IMCOM64595.2025.10857551.

[11] A. A. Firdaus and A. I. Hadiana, "Klasifikasi Sentimen pada Aplikasi Shopee Menggunakan Fitur Bag of Word dan Algoritma Random Forest," *Ranah Res. J. Multidiscip. Res. Dev.*, 2024, doi: 10.38035/rrj.v6i5.994.

[12] A. Susanto and I. A. Dzulkarnain, "Analisis Sentimen Data Twitter Topik Ekonomi Dan Industri Dengan Metode Naive Bayes Dan Random Forest," *J. Ilm. Wahana Pendidik.*, 2023, doi: https://doi.org/10.5281/zenodo.8398895.

[13] M. I. Fikri, T. S. Sabrila, and Y. Azhar, "Perbandingan metode naïve bayes dan support vector machine pada analisis sentimen twitter," *SMATIKA J. STIKI Inform. J.*, 2020, doi: 10.32664/smatika.v10i02.455.

[14] T. Sabri, O. El Beggar, and M. Kissi, "Comparative study of Arabic text classification using feature vectorization methods," *Procedia Comput. Sci.*, 2022, doi: 10.1016/j.procs.2021.12.239.

[15] D. Sugiarto, E. Utami, and A. Yaqin, "Perbandingan Kinerja Model TF-IDF dan BOW untuk Klasifikasi Opini Publik Tentang Kebijakan BLT Minyak Goreng," *J. Tek. Ind.*, 2022, doi: 10.25105/jti.v12i3.15669.

[16] A. P. Pimpalkar and R. J. R. Raj, "Influence of pre-processing strategies on the performance of ML classifiers exploiting TF-IDF and BOW features," *ADCAIJ Adv. Distrib. Comput. Artif. Intell. J.*, 2020, doi: 10.14201/adcaij2020924968.

[17] A. N. Dzulhijjah, S. Anraeni, and S. Sugiarti, "Klasifikasi kematangan citra labu siam menggunakan metode KNN (K-Nearest Neighbor) dengan ekstraksi fitur HSV (Hue, Saturation, Value)," *Bul. Sist. Inf. dan Teknol. Islam*, 2021, doi: 10.33096/busiti.v2i2.808.

[18] H. Darwis, N. Wanaspati, and S. Anraeni, "Support Vector Machine untuk Analisis Sentimen

Masyarakat Terhadap Penggunaan Antibiotik di Indonesia," *Indones. J. Comput. Sci.*, 2023, doi: https://doi.org/10.33022/ijcs.v12i4.3320.
[19] W. T. H. Putri and R. Hendrowati, "Penggalian Teks Dengan Model Bag of Words Terhadap Data Twitter," 2018, *academia.edu*.
[20] D. Septiani and I. Isabela, "Analisis term frequency inverse document frequency (tf-idf) dalam temu kembali informasi pada dokumen teks," 2022.
[21] V. Analytics, "Term frequency inverse document frequency (TF-IDF)," 2023.
[22] N. A. Supriadi, A. R. Manga, R. Adawiyah, and ..., "Application of Ensemble Machine Learning for DDoS Detection in Complex Network Environments," *Proc. 2025 19th Int. Conf. Ubiquitous Inf. Manag. Commun. IMCOM 2025*, 2025, doi: 10.1109/IMCOM64595.2025.10857516.