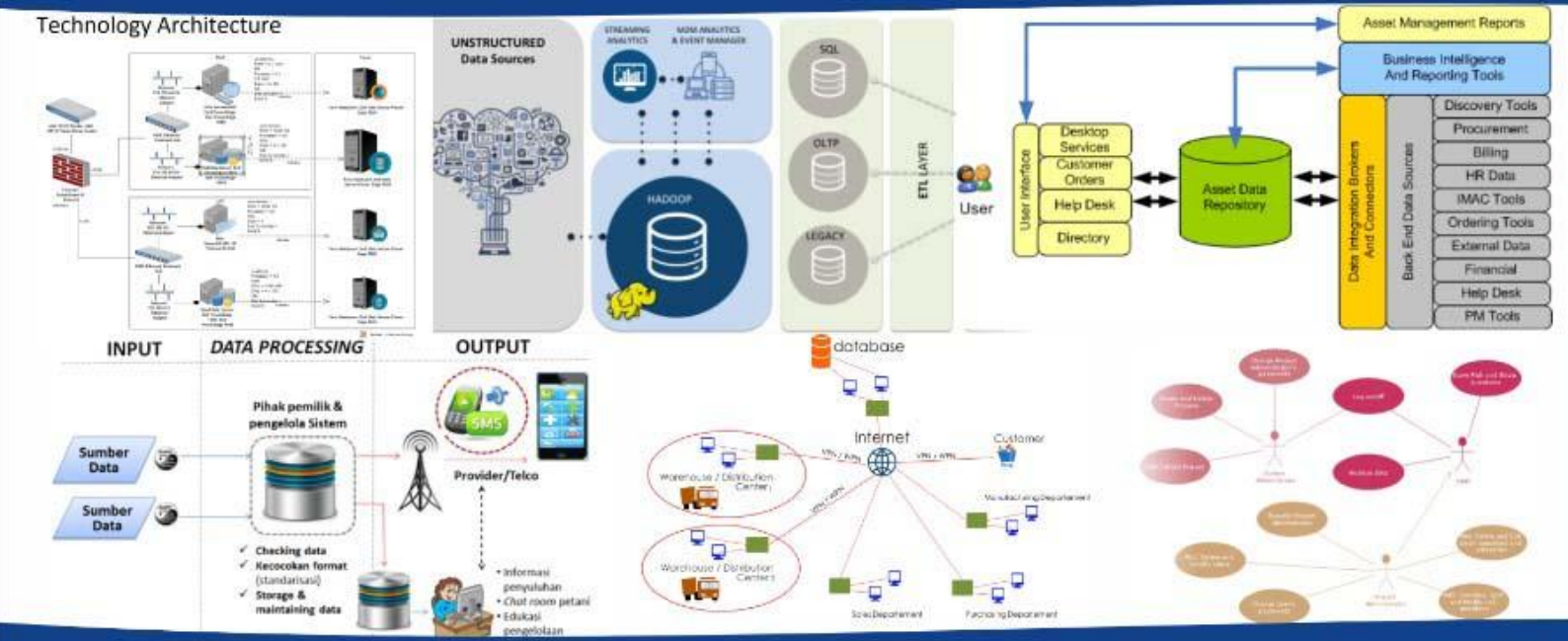
ISSN : 2614-1701 *(Cetak)*
ISSN : 2614-3739 *(Online)*

Volume 5, Nomor 1, Januari 2022

MISI

JURNAL MANAJEMEN INFORMATIKA & SISTEM INFORMASI



9 772614 170004

Diterbitkan Oleh LPPM STMIK Lombok

Jln. Basuki Rahmat No.105 Praya, Lombok Tengah - NTB
Telp dan Fax (0370) 654310 - e-journal.stmiklombok.ac.id/jsi
email. lppm@stmiklombok.ac.id

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 5, Nomor 1, Januari 2022.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI 10.36595/misi.v5i1
http://e-journal.stmiklombok.ac.id/index.php/misi

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 5, Nomor 1, Januari 2022.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI 10.36595/misi.v5i1
http://e-journal.stmiklombok.ac.id/index.php/misi

DAFTAR ISI

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 5, Nomor 1, Januari 2022.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI 10.36595/misi.v5i1
http://e-journal.stmiklombok.ac.id/index.php/misi

RANGKING INDEKS BERITA LARANGAN MUDIK PADA PORTAL MEDIA *ONLINE*DENGAN METODE TF-IDF DAN *COSINE SIMILARITY* MENGGUNAKAN *MACHINE LEARNING*

Muhammad Syahrani[1], Kusnadi [2], Bambang Joko Triwibowo[3], Yusuf Arif Setiawan[4], Fariszal Nova Arviantino[5], Didi Rosiyadi[6]

[1,2,3,4,5,6]Program Studi Ilmu Komputer, Universitas Nusa Mandiri, Jakarta, Indonesia

email : 14002621@nusamandiri.ac.id[1]., kusnusamandiri@gmail.com[2]., 14002607@nusamandiri.ac.id[3]., 14002634@nusamandiri.ac.id[4]., 14002635@nusamandiri.ac.id[5]., didi.rosiyadi@gmail.com[6].

ABSTRACT

The efforts of the Indonesian government in preventing the spread of the Covid 19 virus with the issuance of regulations applied to the regional level. And the annual tradition of Indonesian people homecoming Lebaran 2021 has been banned. News opinions about banning lebaran homecoming both in print and online media and in social media are also widely discussed, of course, people who will go home feel confused with the news and do not know when and when it is enforced.This researcher experimented to collect news that is on online media portals.The collection of news is used as a dataset, then preprocessing includes the stages of tokenizing, filtering and stemming. Accurate search for news information can use the vector space model algorithm by calculating IDF TF and cosine similarity in each news title (document) and in this paper researchers using machine learning.The dataset used 5 news headlines that were labeled D1, D2, D3, D4, and D5, respectively. The results showed that the highest ranking of homecoming ban news indexes was found in document 5(D5) with a score of 0.612.These results strengthen the purpose of the study is to find out the appropriate keywords to be used in order to obtain relevant news and in accordance with the wishes by calculating and calculating the results of cosine similarity values.

Keywords : data mining, retrieval information, vector space model, TF-IDF, cosine similarity

ABSTRAK

Usaha pemerintah Indonesia dalam pencegahan penyebaran virus *Covid* 19 dengan dikeluarkannya peraturan yang diterapkan sampai tingkat daerah. Dan tradisi tahunan masyarakat Indonesia mudik lebaran 2021 telah dilarang. Opini berita tentang pelarangan mudik lebaran baik dimedia cetak maupun media online dan dimedia sosialpun ramai diperbincangkan, tentu masyarakat yang akan mudik merasakan kebingungan dengan pemberitaan tersebut dan belum mengetahui kapan dan sampai kapan diberlakukan. Hal ini peneliti bereksperimen mengumpulkan berita-berita yang ada di portal media *online*. Kumpulan berita tersebut dijadikan dataset, selanjutnya dilakukan *preprocessing* meliputi tahapan *tokenizing, filtering* dan *stemming*. Pencarian informasi berita yang akurasi dapat menggunakan algoritma *vector space model* dengan menghitung TF IDF dan *cosine similarity* pada setiap judul berita (dokumen) dan pada paper ini peneliti dengan menggunakan *machine learning*. Dataset yang digunakan 5 judul berita yang masing-masing diberi label D1, D2, D3, D4, dan D5. Hasil penelitian menunjukan bahwa rangking indek berita larangan mudik yang paling tinggi terdapat pada dokumen 5(D5) dengan skor 0,612. Hasil tersebut menguatkan akan tujuan penelitian yaitu untuk mengetahui *keyword* yang cocok digunakan agar dapat memperoleh berita yang relevan dan sesuai keinginan dengan menghitung dan merangking hasil nilai *cosine similarity.*

Kata kunci : data mining, informasi *retrieval*, *vector space model, TF-IDF, cosine similarity*

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 5, Nomor 1, Januari 2022.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI 10.36595/misi.v5i1
http://e-journal.stmiklombok.ac.id/index.php/misi

1. PENDAHULUAN

Wabah virus corona sedang melanda di seluruh negara yang mengakibatkan terhambatnya hampir disemua sektor kehidupan. Usaha semua negara termasuk di Indonesia, pemerintah Indonesia berupaya untuk pencegahan virus Covid 19 ini dengan di keluarkannya berbagai peraturan dari pusat sampai tingkat daerah agar diberlakukan[1].

Pemerintah Indonesia dalam penekanan lajunya penularan Covid 19 pada masyarakat mengeluarkan kebijakan-kebijakan, seperti penutupan tempat keramaian dan kerumunan yaitu penutupan sementara sekolah, kantor, mall, pasar, tempat rekreasi dan tempat umum lainya[2]. Dan tradisi tahunan masyarakat yang merantau akan mudik lebaran pemerintah mengeluarkan peraturan tentang peniadaan mudik lebaran tahun 2021[3].

Larangan mudik lebaran tahun 2021 kembali dibahas untuk menekan bertambahnya jumlah kasus covid 19 pasca mudik. Menurut Dirjen Kementerian Budi Setiadi, transportasi umum dan pribadi yang bergerak antar daerah dalam rangka mudik lebaran akan dilarang beroperasi [4].

Opini pada masyarakat tentang pelarangan mudik lebaran 2021 diberbagai media baik media online, media cetak maupun media elektronik ramai dibicarakan. Banyaknya pemberitaan tersebut berakibat problem mulai tampak jika pencarian berita tidak sesuai dengan harapan. Berarti berita yang didapat kurang akurat dan tidak bermakna dengan kemunculan berita yang diinginkan dan kyewords yang dipakai kurang berhasil.

Pada saat pencarian suatu informasi pada web yang seringkali informasi pentig terlewatkan sedangkan yang kurang bermakna justru mudah didapatkan[1]. Pemberitaan larangan mudik lebaran 2021 ini, peneliti mengusulkan kajian dengan algoritma information retrieval sebagai pemprosesan indeks berita teks dokumen. Tujuan penelitian yaitu untuk mengetahui keywords yang cocok digunakan agar dapat memperoleh berita yang relevan dan sesuai keinginan dengan menghitung dan merangking hasil nilai Cosine Similarity. Dataset yang dipakai pada penelitian ini yaitu berupa dokumen teks berita tentang virus corona dan pelarangan mudik yang memiliki kyewords yang sama.

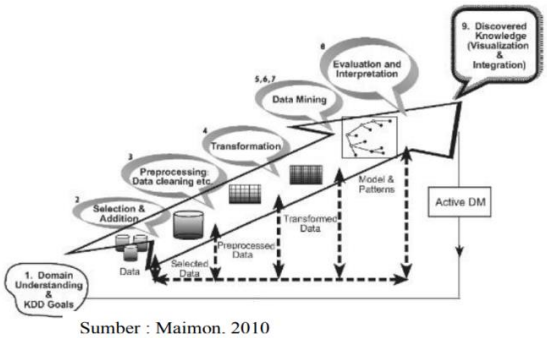
2. TINJAUAN PUSTAKA DAN TEORI

2.1. Data Mining

Data mining adalah ilmu pengetahuan tentang teknologi yang universal dan banyak diterapkan pada cabang disiplin ilmu, seperti ilmu statistik, matematika, komputer, teknik, biologi, fisika, ekonomi dan lainnya[1]. Data mining dapat diartikan sebagai tahap pencarian pola pada data dan berdasarkan fungsi data mining mengelompokkan menjadi deskripsi, estimasi, prediksi, klasifikasi, klasterisasi, danasosiasi[5].

Data mining bisa juga sebagai ilmu yang dapat menggali *dataset* cukup besar untuk memperoleh informasi pengetahuan dan awalnya belum dikenal serta memiliki potensi manfaat yang banyak bagi kehidupan. Teks tambang yaitu suatu bidang khusus dari data tambang dan bisa didefinisikan juga sebagai suatu proses penggalian informasi yang mana seorang *user* melakukan interaksi terhadap kumpulan dokumen dengan memakai alat bantu analisis[6],[5].

Tidak sedikit yang memakai data mining sebagai istilah terkenal dan menjadi utama pada tahap penemuan pengetahuan dalam basis data (KDD).KDD yaitu suatu tahapan yang rapi untuk memperkenalkan pola yang akurat, memiki manfaat dan bisa dipahami dari dataset yang kompleks serta berjumlah banyak[1].



Sumber : Maimon. 2010

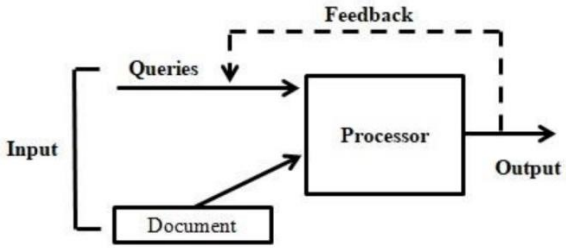
Gambar 1. Proses KDD

2.2. Informasi *Retrieval*

Informasi retrieval merupakan metode untuk memperoleh kembali informasi baru pada dokumen yang sesuai dengan keinginan pengguna dengan terlebih dulu melakukanperangkingan. Rangking adalah bagian dari cabang ilmu temu kembali informasi serta bagian terpenting dari problem mencari informasi, seperti penyaringan informasi, misalnya pengambilan dokumen, menyaring

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 5, Nomor 1, Januari 2022.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI 10.36595/misi.v5i1
http://e-journal.stmiklombok.ac.id/index.php/misi

informasi dan menempatkan iklan *online*, dan sebagainya[7].



Sumber : C.J. Rijsbergen, 1979

Gambar 2. Diagram Alur Informasi *Reterieval*

Pencarian informasi bertujuan untuk memperoleh semua dokumen yang relevan dan dengan waktu yang sama mendapatkan seminim mungkin dokumen yang tidak relevan[8].

2.3. *Vector Space Model*

Vector Space Model(VSM)merupakanproses algoritma yang merubah teks yang panjang dan berbeda, seperti kalimat, paragraf atau dokumen menjadi vektor *numerik* untuk dimasukkan ke dalam aplikasi *down-stream* (seperti deteksi kesamaan atau algoritma *machine learning*)[9]. TF IDF yaitu metode vektorisasi teks yang paling dasar, mendefinisikan ruang di mana setiap istilah dalam kosa kata diwakili oleh dimensi yang terpisah dan orthogonal[10].VSM adalah metode untuk mengetahui tingkat kemiripan kata dengan cara memberikan bobot pada kata dengan memakai metode pembobotan TF-IDF. Dokumen dan *kyewords* dijadikan sebagai sebuah vektor yang mempunyai arah dan jarak. Hubungan tiap dokumen ketiap *kyewords* yang berdasarkan pada *similarity* antara vektor dokumen dan vektor *kyeword*[8]. Sistem bekerja mempunyai dua tahapan yakni melaksanakan tahapan preprosessing terhadap basis data serta melaksanakan penerapan metode tertentu untuk memperoleh nilai similaritas antara dokumen didalam basis data yang sudah dipreprosessing dengan kueri pemakai [11] dan pemberian pembobotan lokal disetiap dokumen telah selesai sesuai yang diinginkan.

2.4. *Term frequency inverse document frequency* (TF-IDF)

Menurut beberapa peneliti, *term trequency inverse document frequensi*(TF-IDF) adalah salah satu cara vektorisasi yang umum buat data tekstual dengan jumlah banyak dan bervariasi serta setiap istilah dianggap sebagai dimensi berbeda secara ortogonal pada dimensi lainnya[10], [12]. Peneliti lainnya mendefinisikan metode TF-IDF yaitu metode perhitungan mencari nilai bobot suatu kata terhadap dokumen dan dengan hasil yang lebih baik[13] dan terkenal efisiensi, mudah dan hasilnya yang akurat[14]. Juga, ada peneliti yang mengartikan algoritma TF-IDF adalah teknik pemberian bobot yang berbasis statistik [15], pada tiap-tiap kata dalam kalimat akan diberi bobot kemudian dilakukan pengurutan sesuai nilai bobot pada masing-masing kalimat tersebut[16], [11].

Dan dalam dokumen, frekuensi istilah (TF) mengacu pada berapa kali kata muncul dalam dokumen. *Inverse document frequency* (IDF) adalah ukuran kepentingan umum dari sebuah kata. Sebuah kata TF-IDF adalah nilai yang diperoleh dengan mengalikan TF dan IDF. Semakin besar TF-IDF dari sebuah kata, semakin penting kata itu dalam dokumen[9]. Algoritma TF-IDF ialah metode yang dipergunakan untuk mencari nilai bobot dari setiap dokumen yang terkait dengan *kyewords* yang dipergunkan [17]. TF-IDF jika didefinisikan masing-masing yakni, TF (*Term Frequesy*) yaitu banyaknya kata yang muncul didalam setiap dokumen. Sedangkan IDF *(Inverse Dokumen Frequecy)* yaitu perhitungan jumlah frekuensi pada sekumpulan dokumen[15].

TF-IDF bisa dirumuskan sebagai berikut:

$$idf = \log(D/df) \quad (1)$$

Keterangan:
D = Jumlah dokumen
df = frekuensi kata pada dokumen
Rumus perehitungan pembobotan terdapat pada persamaaan berikut:

$$Wd = tfdt * ID \quad (2)$$

Keterangan:
d = dokumen ke-d dari dokumen yang ada di basis data
t = kata ke-t dari *kyewords*
tf = jumlah kata yang dicari pada dokumen
W = bobot dokumen ke-d terhadap *kyewords* ke-t

2.5. *Cosine Similarity*

Perhitungan kesamaan yaitu menghitung kesamaan masing-masing pasangan paten berdasarkan *sim cosine*. vektor yang sesuai dari ruang vektor yang berbeda[18]. Untuk

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 5, Nomor 1, Januari 2022.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI 10.36595/misi.v5i1
http://e-journal.stmiklombok.ac.id/index.php/misi

membandingkan pasangan paten dari model TF IDF tambahan, sebuah studi baru-baru ini mengusulkan untuk mengganti IDF dari paten yang terakhir dengan paten sebelumnya sehingga kedua vektor memiliki skala waktu yang sama untuk dibandingkan[19]. Perhitungan jarak selain menggunakan *euclidean distance* juga dapat menggunakan metode *cosine similarity*[20].

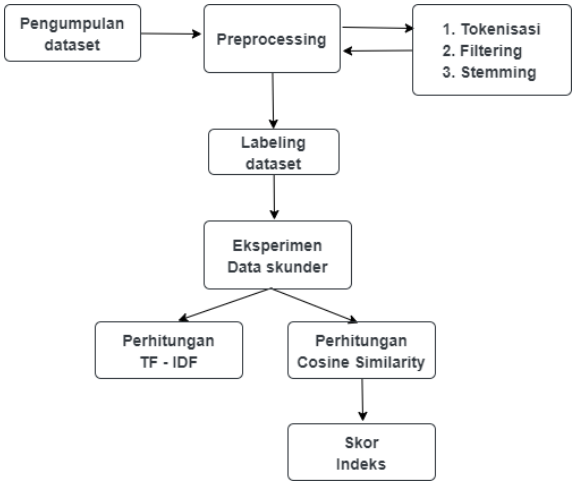
Metode untuk menghitung tingkat kemiripan antara dua buah objek[14], misalnya D1 dan D2 dijelaskan pada dua buah vektor dengan memakai *kyewords* dari sebuah dokumen sebagai ukuran yang juga disebut *cosine similarity*[21].

Cosine similarity jika dirumus dalam pesamaan sebagai berikut:

$$Cos\ \alpha = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \qquad (3)$$

Keterangan :
A = Vektor A, yang akan dibandingkan kemiripan
B = Vektor B, yang akan dibandingkan kemiripan
A.B = dot *product* antara vektor A dan vektor B
|A| = Panjang vektor A
|B| = Panjang vektor B
|A||B| = *cross product* antara panjang vektor A dan panjang vektor B

3. METODE PENELITIAN



Gambar 3. *Flowmap* penelitian

Penelitian ini memakai metode eksperimen dengan menggunakan algoritma TF IDF dan *consine similarity* untuk menghitung rangking indeks dari dokumen teks yang diperoleh dari berita *online* tentang pelarangan mudik dan covid 19. Berdasarkan Gambar 3. *Flowmap*penelitian ini, dapat diuraikan sesuai tahapan sebagaimana berikut:

3.1. Tahapan pengumpulan data

Pada penelitian, hal yang sangat penting adalah data, dengan banyaknya data dan bervariasi akan berpengaruh pada hasil penelitian. Dalam mengumpulkan data dengancara literasi yaitu cara mengumpulkan datanya berupa data sekunder dari kumpulan berita yang beredar di portal media *online* liputan6.com, suarasurabaya.net, cnbcindonesia.com, dan tirto.id.

3.2. Tahapan *preprocessing*

Tahap persiapan data adalah tahap proses menyiapkan data yang memiliki tujuan untuk memperoleh data yang benar dan siap diolah dalam analisis penelitian. Dalam penambangan teks, proses pertama yang akan dilaksanakan ialah tahapan pra pemrosesan[14]. *Preprocessing* merupakan tahap proses yang mempunyai tujuan untuk membuat sekumpulan data bersumber dari database sekunder yang tidak beraturan hinggga dataset siap untuk diolah.

Tahan tersebut tebagi menjadi 3, yaitu:

A. Tokenisasi

Tokenisasi yaitu proses pengenalan dokumen dan melakukan pemisahan kalimat menjadi kata tunggal. Selanjutnya melakukan proses penghilangan karakter spesial, misalnya tanda baca dan dirubah semua huruf jadi huruf kecil. Tokenisasi mempunyai sasaran untuk memperoleh kata yang unik dari tiap dokumen[9], maka dapat diperoleh nilai frekuensi dan nilai bobot disetiap kata yang terdapat pada tiap dokumen[22].

Contok tokenisasi:
"Jadwal Pelarangan Mudik Lebaran dan Aturan Baru"

Proses tokenisasi:
1. Menghilangkan *special character*

Hasil = Jadwal Pelarangan Mudik Lebaran dan Aturan Baru
2. Memisahkan menjadi kata tunggal dan unik

Hasil = Jadwal larang mudik Lebar dan Atur Baru
3. Merubah menjadi huruf kecil

Hasil = jadwal larang mudik lebar dan atur baru

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 5, Nomor 1, Januari 2022.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI 10.36595/misi.v5i1
http://e-journal.stmiklombok.ac.id/index.php/misi

B. *Filtering*

Tahapan ini melakukan penghapusan kata penghubung dan kata sering muncul. Hasil tokensasi tersebut dilanjutkan melaksanakan *stopword*, yaitu kata "dan", sehingga hasil filtering menjadi: jadwal larang mudik lebar atur baru

C. *Stemming*

Algoritma yang dipakai ialah stemmer porter, yaitu sebuah tahapan mencari kata dasar. Pada bahasa Indonesia sering memakai kaedah awalan + kata dasar + akhiran[20]. Maka hasil akhirnya dari *preprocessing* pada contok tersebut yaitu kata-kata: jadwal, larang, mudik, lebar, atur, baru

3.3. Tahapan perhitungan TF-IDF dan *cosine similarity*

Dataset sekunder setelah dilakukan tahap *preprocessing* yaitu meliputi tokenisasi, *filtering* dan *stemming,* kemudian melakukan perhitungan nilai TF-IDF dan *cosine similarity* dengan alat bantu *machine learnig* Anacoda-Yuphiter.

4. HASIL DAN PEMBAHASAN

4.1. Dataset sekunder

Dataset sekunder diperoleh dari judul berita online dari beberapa sumber yang berkenaan dengan topik "peraturan larangan mudik 2021", selanjutnya melakukan *preprocessing* yaitu meliputi tokenisasi, filtering dan stemming diperoleh data teks sebagai berikut:

D1= siasat orang mudik lebih awal larang mudik
D2= simak atur kecuali jalan lama larang mudik
D3= pemerintah jelas alas larang mudik lebar
D4= larang mudik sudah final sektor usaha nangis
D5= jadwal larang mudik lebar atur baru
Q= atur pemerintah larang mudik

4.2. Perhitungan TF-IDF dokumen dan kueri

Hasil *prossing machine laerning*

```
#perhitungan TF-IDF Dokumen
corpus
['siasat orang mudik lebih awal larang
mudik',
 'simak atur kecuali jalan lama larang
mudik',
 'pemerintah jelas alas larang mudik
lebar',
 'larang mudik sudah final sektor usaha
nangis',
 'jadwal larang mudik lebar atur baru']
print(response)
  (0, 10)     0.21027393278978984
  (0, 2)      0.4412834593392252
  (0, 12)     0.4412834593392252
  (0, 13)     0.4205478655795797
  (0, 15)     0.4412834593392252
  (0, 18)     0.4412834593392252
  (1, 9)      0.44258920496410614
  (1, 6)      0.44258920496410614
  (1, 8)      0.44258920496410614
  (1, 1)      0.35707818379679507
  (1, 19)     0.44258920496410614
  (1, 10)     0.21089612757628357
  (1, 13)     0.21089612757628357
  (2, 11)     0.39820278266020154
  (2, 0)      0.4935620852501244
  (2, 7)      0.4935620852501244
  (2, 16)     0.4935620852501244
  (2, 10)     0.23518497814732847
  (2, 13)     0.23518497814732847
  (3, 14)     0.42819132662403886
  (3, 21)     0.42819132662403886
  (3, 17)     0.42819132662403886
  (3, 4)      0.42819132662403886
  (3, 20)     0.42819132662403886
  (3, 10)     0.20403546140282616
  (3, 13)     0.20403546140282616
  (4, 3)      0.5159888056221924
  (4, 5)      0.5159888056221924
  (4, 11)     0.4162965194462718
  (4, 1)      0.4162965194462718
  (4, 10)     0.24587143056789498
  (4, 13)     0.24587143056789498

>>> from
sklearn.feature_extraction.text import
TfidfVectorizer

>>> vectorizer =
TfidfVectorizer(stop_words='english')
>>> response =
vectorizer.fit_transform(corpus)
print(response)

Vectorizer get feature names
['alas',
 'atur',
 'awal',
 'baru',
 'final',
 'jadwal',
 'jalan',
 'jelas',
 'kecuali',
 'lama',
 'larang',
 'lebar',
 'lebih',
 'mudik',
 'nangis',
 'orang',
 'pemerintah',
 'sektor',
```

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 5, Nomor 1, Januari 2022.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI 10.36595/misi.v5i1
http://e-journal.stmiklombok.ac.id/index.php/misi

```
 'siasat',
 'simak',
'sudah',
 'usaha'

Import pandas as pd
```

```
Df = pd.DataFrame(response.todense().T,
  index =
vectorizer.get_feature_names(),
columns = [f'D{i+1}'for i in
range(len(corpus))])
df
```

Term (t)	D1	D2	D3	D4	D5
alas	0.000000	0.000000	0.493562	0.000000	0.000000
atur	0.000000	0.357078	0.000000	0.000000	0.416297
awal	0.441283	0.000000	0.000000	0.000000	0.000000
baru	0.000000	0.000000	0.000000	0.000000	0.515989
final	0.000000	0.000000	0.000000	0.428191	0.000000
jadwal	0.000000	0.000000	0.000000	0.000000	0.515989
jalan	0.000000	0.442589	0.000000	0.000000	0.000000
jelas	0.000000	0.000000	0.493562	0.000000	0.000000
kecuali	0.000000	0.442589	0.000000	0.000000	0.000000
lama	0.000000	0.442589	0.000000	0.000000	0.000000
larang	0.210274	0.210896	0.235185	0.204035	0.245871
lebar	0.000000	0.000000	0.398203	0.000000	0.416297
lebih	0.441283	0.000000	0.000000	0.000000	0.000000
mudik	0.420548	0.210896	0.235185	0.204035	0.245871
nangis	0.000000	0.000000	0.000000	0.428191	0.000000
orang	0.441283	0.000000	0.000000	0.000000	0.000000
satgas	0.000000	0.000000	0.493562	0.000000	0.000000
sektor	0.000000	0.000000	0.000000	0.428191	0.000000
siasat	0.441283	0.000000	0.000000	0.000000	0.000000
simak	0.000000	0.442589	0.000000	0.000000	0.000000
sudah	0.000000	0.000000	0.000000	0.428191	0.000000
usaha	0.000000	0.000000	0.000000	0.428191	0.000000

```
#Perhitungan TF-IDF Queri
corpus
['atur larang mudik']
print(response
  (0, 2)      0.5
  (0, 1)          0.5
  (0, 3)          0.5
  (0, 0)          0.5
```

```
df
```

Term (t)	Q
atur	0.5
larang	0.5
mudik	0.5
satgas	0.5

```
Vectorizer get feature names
Vectorizer get feature names
['atur', 'larang', 'mudik', 'satgas']

Import pandas as pd
Df=pd.DataFrame(response.todense().T,
index=vectorizer.get_feature_names(),
columns    =    [f'Q'for    i    in    range
    (len(corpus))])
```

4.3. Perhitungan *cosine similarity*

Hasil perhitungan similaritas dengan *machine learning* diperoleh sebagai berikut:

```
#Mencari Consine Similarity
from    sklearn.feature_extraction.text
import CountVectorizer
```

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 5, Nomor 1, Januari 2022.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI 10.36595/misi.v5i1
http://e-journal.stmiklombok.ac.id/index.php/misi

```
from sklearn.metrics.pairwise import cosine_similarity

query = "atur pemerintah larang mudik"
d1 = "siasat orang mudik lebih awal larang mudik"
d2 = "simak atur kecuali jalan lama larang mudik"
d3 = "pemerintah jelas alas larang mudik lebar"
d4 = "larang mudik sudah final sektor usaha nangis"
d5 = "jadwal larang mudik lebar atur baru"

Kumpulanteks = [query,d1,d2,d3,d4,d5]
X = CountVectorizer(encoding='latin-1', binary=False)
#x_hentihapus
=CounVectorizer(encoding='latin-1', binary=False, stop_word='english')
#untuk filter kata penghubung, awalan dan akhiran supaya hilang
vector = X.fit_transform(Kumpulanteks)
cosim = cosine_similarity(vector[0], vector)
urutan = cosim.argsort()

print("Text Query : ", query)
print("Text D1 : ", d1)
print("Text D2 : ", d2)
print("Text D3 : ", d3)
print("Text D4 : ", d4)
print("Text D5 : ", d5)
print()
print(cosim)

print("Data paling mirip data : d",urutan[0][len(Kumpulanteks)-2])
print("Nilai kemiripan/cosim : ",cosim[0][urutan[0][len(Kumpulanteks)-2]])

print("Data paling mirip data : d",urutan[0][len(Kumpulanteks)-3])
print("Nilai kemiripan/cosim : ",cosim[0][urutan[0][len(Kumpulanteks)-3]])

print("Data paling mirip data : d",urutan[0][len(Kumpulanteks)-4])
print("Nilai kemiripan/cosim : ",cosim[0][urutan[0][len(Kumpulanteks)-4]])

print("Data paling mirip data : d",urutan[0][len(Kumpulanteks)-5])
print("Nilai kemiripan/cosim : ",cosim[0][urutan[0][len(Kumpulanteks)-5]])

print("Data paling mirip data : d",urutan[0][len(Kumpulanteks)-6])
print("Nilai kemiripan/cosim : ",cosim[0][urutan[0][len(Kumpulanteks)-6]])
Text Query : atur pemerintah larang mudik
Text D1 : siasat orang mudik lebih awal larang mudik
Text D2 : simak atur kecuali jalan lama larang mudik
Text D3 : pemerintah jelas alas larang mudik lebar
Text D4 : larang mudik sudah final sektor usaha nangis
Text D5 : jadwal larang mudik lebar atur baru

[[1. 0.5 0.56694671 0.61237244 0.37796447 0.61237244]]
Data paling mirip data : d 5
Nilai kemiripan/cosim : 0.6123724356957947
Data paling mirip data : d 3
Nilai kemiripan/cosim : 0.6123724356957946
Data paling mirip data : d 2
Nilai kemiripan/cosim : 0.5669467095138407
Data paling mirip data : d 1
Nilai kemiripan/cosim : 0.5
Data paling mirip data : d 4
Nilai kemiripan/cosim : 0.3779644730092272
```

Berdasarkan hasil perhitungan *machine learning* dari nilai *cosine similarity* diatas, bahwa 5 judul berita dengan topik seputar larangan mudik yang dijadikan sebagai teks dokumen dengan queri "peraturan pemerintah larangan mudik" didapat rangking indeks sebagai berikut:
Rangking indeks ke 1 terdapat pada dokumen 5 (D5) memiliki skor 0,6123724356957947
Rangking indeks ke 2 terdapat pada dokumen 3 (D3) memiliki skor 0,6123724356957946
Rangking indeks ke 3 terdapat pada dokumen 2 (D2) memiliki skor 0,5669467095138407
Rangking indeks ke 4 terdapat pada dokumen 1 (D1) memiliki skor 0,5
Rangking indeks ke 5 terdapat pada dokumen 4 (D4) memiliki skor 0,3779644730092272
Jadi, rangking indek berita larangan mudik yang paling tinggi pada penelitian ini terdapat pada dokumen 5(D5) dengan skor 0,6123724356957947.

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 5, Nomor 1, Januari 2022.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI 10.36595/misi.v5i1
http://e-journal.stmiklombok.ac.id/index.php/misi

5. KESIMPULAN DAN SARAN

5.1. KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkanbahwametode TF-IDF dan *cosine similarity*menggunakan *machine learning*telah bekerja maksimum dengan menghasilkan rangking indeks berita larangan mudik yang tertinggi pada penelitian ini terdapat pada dokumen 5 (D5) dengan skor 0,612.

5.2. SARAN

Kami sadar pada penelitian terdapat kekurangan pada pra-pemrosesan yaitu pada tahapan tokenisasi, penyaringan dan *stemming* tidak menggunakan alat bantu *machine learning*.

6. UCAPAN TERIMA KASIH

Kami ucapkan terima kasih kepada pimpinan Universitas Nusa Mandiri dan dosen pembimbing yang telah menberikan dukungan secara moril dan bimbingannya, sehingga pembuatan paper ini selesai.

Daftar Pustaka:

[1] N. Suwela, "Ranking Index Berita New Normal dengan Metode Information Retrieval Menggunakan Vector Space Model," *STRING (Satuan Tulisan Ris. dan Inov. Teknol.*, vol. 5, no. 1, p. 61, 2020, doi: 10.30998/string.v5i1.6479.

[2] K. D. N. RI, "Instruksi Mendagri No.17 tahun2021 ttg PPKM Berbasis mikro dan Mengoptimalkan Posko Penanganan Corona Virus Disease 2019 Di Tingkat Desa dan Kelurahan Untuk Pengendalian Penyebaran Corona Virus Disease 2019," pp. 1–19, 2021.

[3] Satuan Tugas Penanganan Covid-19, "Peniadaan Mudik Hari Raya Idul Fitri Tahun 1442 Hijriah Dan Upaya Pengendalian Penyebaran Corona Virus Disease 2019 (Covid-19) Selama Bulan Suci Ramadhan 1442 Hijriah," *Satgas Covid -19*. p. 1, 2021, [Online]. Available: https://covid19.go.id/.

[4] P. T. Dan and R. Logistik, "Analisis Sentimen Twitter tentang Covid-19 Menggunakan Istilah,", Seminar Internasional Teknologi Informasi (ITIS) 2020 Surabaya, Indonesia, 14-16 Oktober 2020, pp. 14–16, 2021.

[5] E. D. Sikumbang, "Penerapan Data Mining Penjualan Sepatu Menggunakan Metode Algoritma Apriori," *J. Tek. Komput. AMIK BSI*, vol. Vol 4, No., no. September, pp. 1–4, 2018.

[6] D. Game and O. Pada, "Vector space model," Prosiding SNST ke-7 Tahun 2016, pp. 73–78, 2016.

[7] A. A. Abdillah and I. B. Muktyas, "Implementasi vector space model untuk pencarian dokumen," no. May 2013, 2015.

[8] F. Amin, "Implementasi Search Engine (Mesin Pencari) Menggunakan Metode Vector Space Model," *J. Ilm. Din. Tek.*, vol. 5, no. 1, pp. 45–58, 2011.

[9] H. Yuan, Y. Tang, W. Sun, and L. Liu, "A detection method for android application security based on TF-IDF and machine learning," *PLoS One*, vol. 15, no. 9 September, pp. 1–19, 2020, doi: 10.1371/journal.pone.0238694.

[10] O. Shahmirzadi, A. Lugowski, and K. Younge, "Text Similarity in Vector Space Models :," pp. 659–666, 2019, doi: 10.1109/ICMLA.2019.00120.

[11] Y. Alkhalifi, W. Gata, A. Prasetyo, and I. Budiawan, "Analisis Sentimen Penghapusan Ujian Nasional pada Twitter Menggunakan Support Vector Machine dan Naïve Bayes berbasis Particle Swarm Optimization," vol. 6, no. 2, pp. 71–78, 2020.

[12] D. Sari, "Latent Semantic Indexing for Indonesian Text Similarity, International Journal of Engineering & Technology, pp.1-6, vol.7.2018"

[13] J. Wang, "Text Similarity Calculation Method Based on Hybrid Model of LDA and TF-IDF," pp. 1–8, 2019.

[14] A. Riyani, M. Zidny, and A. Burhanuddin, "Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen," vol. 2, no. 1, pp. 23–27, 2019.

[15] S. Wahyunita, Y. Azhar, and N. Hayatin, "Analisa Sentimen Tweet Berbahasa Indonesia dengan Menggunakan Metode Pembobotan Hybrid TF-IDF pada Topik Transportasi Online," *J. Repos.*, vol. 2, no. 2, p. 185, 2020, doi: 10.22219/repositor.v2i2.238.

[16] G. Vinodhini and R. Chandrasekaran, "Sentiment Analysis and Opinion Mining : A Survey International Journal of

Advanced Research in Sentiment Analysis and Opinion Mining : A Survey," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 2, no. 6, pp. 283–292, 2012.

[17] M. Nurjannah and I. Fitri Astuti, "Penerapan Algoritma Term Frequency-Inverse Document Frequency (TF-IDF) Untuk Text Mining Mahasiswa S1 Program Studi Ilmu Komputer FMIPA Universitas Mulawarman Dosen Program Studi Ilmu Komputer FMIPA Universitas Mulawarman," *J. Inform. Mulawarman*, vol. 8, no. 3, pp. 110–113, 2013.

[18] A. Wibawa and U. Pujianto, "Cosine Similarity for Title and Abstract of Economic Journal Classification," no. July, 2020, doi: 10.1109/ICSITech46713.2019.8987547.

[19] B. Kelly and M. Taddy, "Measuring Technological Innovation over the Long Run ∗," no. January, 2020.

[20] G. Karyono, F. S. Utomo, A. Sistem, and T. Balik, "Temu Balik Informasi Pada Dokumen Teks Berbahasa Indonesia Dengan Metode Vector Space Retrieval Model," *Semin. Nas. Teknol. Inf. dan Terap. 2012*, vol. 2012, no. Semantik, pp. 282–289, 2012.

[21] D. Susandi and U. Sholahudin, "Pemanfaatan Vector Space Model pada Penerapan Algoritma Nazief Adriani , KNN dan Fungsi Similarity Cosine untuk Pembobotan IDF dan WIDF pada Prototipe Sistem Klasifikasi Teks Bahasa Indonesia," *J. ProTekInfo*, vol. 3, no. 1, pp. 22–29, 2016.

[22] H. Kim and J. Baek, "applied sciences Optimization of Associative Knowledge Graph Using TF-IDF Based Ranking Score," 2020.

[23] Imtihan, K., & Fahmi, H. (2020). Analisis Dan Perancangan Sistem Informasi Daerah Rawan Kecelakaan Dengan Menggunakan Geographic Information Systems (GIS). Jurnal Manajemen Informatika dan Sistem Informasi, 3(1), 16-23.

[24] Imtihan, K., & Basri, M. H. (2019). Sistem Informasi Pembuatan Manifest Muatan Kapal Berbasis Dekstop Dan Android. Jurnal Manajemen Informatika dan Sistem Informasi, 2(2), 69-76.

[25] Ashari, M., Zaen, M. T. A., Putri, J. A., Imtihan, K., & Bagye, W. (2022). Prototype Sterilisasi Virus Barang Belanjaan Online Berbasis Arduino. *Jurnal Media Informatika Budidarma*, *6*(1), 120-127.