Jurnal Ilmiah Informatika (JIF)
| ISSN (Print) 2337-8379 | ISSN (Online) 2615-1049 |
Jurnal online di akses di http://ejournal.upbatam.ac.id/index.php/jif

PENGGUNAAN *FITUR WORDCLOUD* DAN *DOCUMENT TERM MATRIX* DALAM *TEXT MINING*

Musthofa Galih Pradana

Universitas Alma Ata, Jl. Brawijaya 99, Yogyakarta 55183, Indonesia

INFORMASI ARTIKEL

Sejarah Artikel
Diterima Redaksi: 26 Februari 2020
Revisi Akhir: 10 Maret 2020
Diterbitkan *Online*: 25 Maret 2020

KATA KUNCI

Data, Text Mining, Term Document Matrix, Word Cloud, Social Media.

KORESPONDENSI

E-mail: mgalihprada@uaa.ac.id

A B S T R A C T

Much information and data can be extracted from social media differences, with more and more social media users. Data in 2019 states that there are 150 million users of social media in Indonesia. Based on the number of active users of social media, it can be exploited for deeper information extraction and analysis. One way that can be done is by taking comment data on social media for further processing or mining. In this research, we do data crawling and utilize the Term Document Matrix and Word Cloud features to find the most frequently written words on Facebook and Twitter social media. The words that appear most often based on the Word Cloud feature will be analyzed to infer from words written on social media. In this study the word that often appears on Facebook is the word garuda for 3621 words and on Twitter is the Indonesian word for 1572. On the Facebook platform the resulting word has a positive tendency because the topics discussed are still around airlines, while on Twitter it has a negative tendency because of the word what appears is a personal name that has a negative tendency for the company.

1. PENDAHULUAN

Dilansir *We Are Social* pada tahun 2019 menyebutkan bahwa pengguna sosial media di Indonesia sebanyak 150 juta pengguna dari 268 juta penduduk Indonesia, atau dalam prosentase sebesar 56% menggunakan internet dalam kehidupan sehari-hari. Ini menandakan bahwa pengguna internet Indonesia tergolong besar. Melihat lebih jauh tentang pengguna internet di Indonesia diperinci lagi ke dalam 150 juta pengguna aktif sosial media, 130 juta pengguna mobile sosial media. Berdasarkan data tersebut, ada banyak data atau opini penduduk Indonesia yang diluapkan di sosial media, dari sekian banyak jumlah pengguna aktif, aka nada banyak sekali data yang dapat dimanfaatkan untuk membuat informasi yang baru.

Cara yang digunakan untuk mengolah data dari sosial media menjadi informasi yang berguna adalah dengan dilakukan pengolahan data atau yang biasa disebut *data mining*. *Data mining* merupakan proses menggali nilai dari sebuah data yang sudah ada secara otomatis [1]. Banyak yang dapat dilakukan dengan *data mining*, teknik yang diterapkan seperti klustering, klasifikasi maupun asosiasi. Mengolah data berupa *text* atau sering disebut dengan *text mining* akan lebih efektif ketika pengelolaan dan pengolahan yang dilakukan menggunakan bantuan perangkat lunak. Pengolahan data yang dilakukan secara manual baik dari *filtering* data, pengelompokan data dan jenis-jenis pengolahan data lainnya akan memakan waktu yang lama.

Penelitian ini memanfaatkan fitur *Word Cloud* yang terdapat pada *software R Studio*. *Word Cloud* salah satu metode untuk memvisualisasikan data teks secara visual. *Word Cloud* populer dalam *text mining* karena mudah dipahami. Dengan menggunakan *word cloud*, gambaran frekuensi kata-kata dapat ditampilkan dalam bentuk yang menarik namun tetap informatif. Ukuran gambar teks dalam *Word Cloud* menyesuaikan dengan frekuensi data, semakin banyak frekuensi kata digunakan, maka semakin besar pula ukuran kata tersebut ditampilkan dalam *Word Cloud*.

Pada penelitian ini juga akan digunakan *Term Document Matrix* untuk menampilkan data kata yang paling sering di tuliskan pengguna sosial media untuk mendapatkan data yang lebih detail dan spesifik. Penggunaan fitur *Term Document Matrix* masih saling terikat dengan penggunaan fitur *Word Cloud*. Kedua fitur ini saling melengkapi dalam penelitian *text mining*.

Objek yang dijadikan tempat pengambilan data adalah sosial media salah satu perusahaan transportasi di Indonesia pada *platform Facebook* dan *Twitter* yang akan dibandingkan masing-masing hasil penggunaan fitur *Word Cloud* dan *Term Document Matrix*.

2. TINJAUAN PUSTAKA

Penelitian mengenai *Text Mining* pernah dilakukan oleh Dedi Dwi Saputra dkk, yang melakukan *text mining* untuk *assignment complaint handling customer*. Penelitian ini menghasilkan kesimpulan *Text Mining* dan model *decision tree* dapat digunakan dalam membuat klasifikasi sebagai dasar dalam pembangunan sistem pendukung keputusan untuk *assignment complaint handling* kepada suatu divisi [2].

Penelitian selanjutnya oleh Musfiroh Nurjannah dkk yang menerapan algoritma *TF IDF* dalam proses *text mining*. Kesimpulan yang dapat diambil dalam penelitian ini adalah Dalam algoritma *TF-IDF*, frekuensi kemunculan sebuah kata *(term)* dalam dokumen tidak mempengaruhi hasil perhitungan bobot dokumen oleh sistem (sifat monotonicity TF-IDF) [3].

Rimbun Siringoringo dkk juga pernah menuliskan tentang penelitian *text mining* dengan mengklasterisasi *sentiment* pada ulasan produk toko online. Kesimpulan dari penelitian ini adalah dengan menerapkan pengujian pada tiga jenis data sentimen yang berbeda diperoleh hasil bahwa *Suppport Vector Machine* dapat bekerja dengan baik pada data ulasan tidak seimbang [4].

Penelitian rujukan yang berikutnya adalah analisis *sentiment* pada maskapai citilink menggunakan metode *Naïve Bayes*. Penelitian yang ditulis oleh Moh Yasid dkk ini menyimpulkan bahwa dengan dataset sebanyak 2000 tweet menghasilkan akurasi yang mencapai 0,778, waktu proses dua menit dua puluh tiga detik dengan kode pemograman PHP [5].

Rujukan berikutnya adalah tulisan Meylita Putri Simatupang dkk yang melakukan *text mining* menggunakan *TF IDF* pada toko Alfamart. Peneitian ini menghasilkan kesimpulan bahwa Algoritma *text mining* dan *term frequency – inverse document frequency (TF-IDF)* mampu melakukan klasifikasi testimoni dengan konotasi positif dan negatif [6].

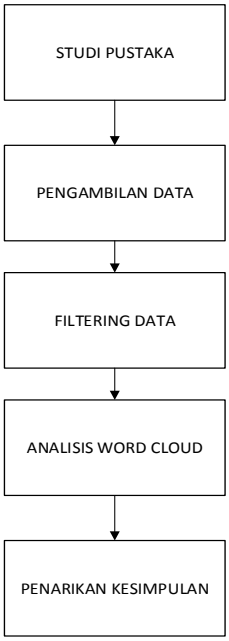
Text mining untuk mencari opini publik yang dilakukan oleh Trya Sovi Kartikasari dkk menyimpulkan bahwa Sistem berhasil mengelompokkan faktor berdasarkan nilai PC yang paling besar dengan teknik analisis faktor yang mengelompokkan kata-kata dalam dimensi yang lebih kecil dan diperlukan masukan jumlah faktor oleh *user* agar dapat memberikan hasil yang maksimal [7].

Penelitian tentang *text mining* juga pernah dilakukan oleh Ni Luh Ratniasih dkk dengan judul Penerapan *Text Mining* dalam *Spam Filtering* untuk Aplikasi Chat. Kesimpulan dari penelitian ini adalah Proses filtering dilakukan dengan tahap text preprocessing dan analyzing sehingga diperoleh kalimat yang dinyatakan sebagai kalimat spam adalah berdasarkan kemunculan kata spam dalam kalimat pesan tersebut, dimana jika nilai W sebuah kalimat pesan lebih besar dari 0.012 *(threshold)* [8].

Penelitian rujukan terkakhir adalah tulisan Firman Abdurrahman dkk dengan menerapkan *TF IDF* untuk penentuan sanksi dalam buku pedoman akademik. Penelitian ini menghasilkan kesimpulan Algoritma *TF-IDF* memiliki tingkat akurasi yang baik, sehingga algoritma *TF-IDF* sangat tepat untuk pencarian suatu *term* [9].

3. METODOLOGI

Adapun langkah atau alur penelitian yang dilakukan dimulai dengan studi pustaka atau mengumpulkan referensi dan rujukan penelitian, kemudian dilakukan pengambilan data dengan teknik *crawling* data di sosial media. Data yang didapat perlu difilter karena pasti banyak data yang kurang memiliki informasi atau makna, sehingga perlu dilakukan *cleaning* atau *filtering* data. Data yang telah difilter kemudian diolah atau dianalisis dengan memanfaatkan fitur *Word Cloud* dan *Term Document Matrix* untuk menghasilkan informasi dan dapat dilakukan penarikan kesimpulan. Adapun detail alur penelitian ditunjukan pada Gambar 1.



Gambar 1. Alur Penelitian

4. HASIL DAN PEMBAHASAN

Pengumpulan Dataset

Dataset atau data penelitian di dapatkan dari sosial media berupa komentar dan kata-kata yang dituliskan penggguna sosial media untuk diketahui informasi yang terkandung di dalamnya. Pengumpulan *dataset* dilakukan dengan teknik *crawling data*. Data yang didapatkan saat *crawling* masih tidak terstruktur dan masih perlu dilakukan pengolahan lebih lanjut. Contoh data yang didapatkan saat *crawling* data ditunjukan pada Tabel 1.

Tabel 1. Data Awal Hasil *Crawling*

Komentar
@chaviation Setelah kami cek, sayangnya promo tersebut telah berakhir di tanggal 20 Februari 2020 kemarin.
Facebook:<post>/comments "Penerbangan jakarta (cgk) ke amsterdam (ams) free wifi juga ya"

Cleaning & Filtering

Dataset komentar masih tidak terstruktur, dan masih banyak *noise* sehingga perlu dilakukan *cleaning* data. Tahap *cleaning* data dilakukan dengan membersihkan atau memilah data sesuai dengan kebutuhan. Langkah pertama pada tahapan *cleaning* data adalah merubah teks ke dalam bentuk *Corpus*. Kemudian, dilakukan pembersihan data, dengan mengganti tanda "/", "@" and "|" menjadi spasi. Contoh penerapan *cleaning* data ditunjukan pada Tabel 2.

Tabel 2. Proses Penghilangan Karakter

Awal	Akhir
RT @IndonesiaGaruda: Kini para penumpang Economy Class dari Jakarta ke Jepang, Korea, Australia & Amsterdam juga sudah bisa #BookYourMeal	RT IndonesiaGaruda: Kini para penumpang Economy Class dari Jakarta ke Jepang, Korea, Australia & Amsterdam juga sudah bisa #BookYourMeal
Facebook:<post>/comments "Penerbangan jakarta (cgk) ke amsterdam (ams) free wifi juga ya"	Facebook:<post> comments Penerbangan jakarta (cgk) ke amsterdam (ams) free wifi juga ya

Langkah selanjutnya dilakukan proses *case folding*, yakni menyeragamkan huruf ke dalam bentuk huruf kecil agar pengkategorian kata menjadi lebih mudah adapun langkahnya ditunjukan pada Tabel 3.

Tabel 3. Proses *Case Folding*

Awal	Akhir
RT IndonesiaGaruda: Kini para penumpang Economy Class dari Jakarta ke Jepang, Korea, Australia & Amsterdam juga sudah bisa #BookYourMeal	rt indonesiagaruda: kini para penumpang economy class dari jakarta ke jepang, korea, australia & amsterdam juga sudah bisa #bookyourmeal
Facebook:<post> comments Penerbangan jakarta (cgk) ke amsterdam (ams) free wifi juga ya	facebook:<post> comments penerbangan jakarta (cgk) ke amsterdam (ams) free wifi juga ya

Langkah berikutnya adalah menghapus tanda baca *(punctuation)* agar benar-benar tinggal teks yang tersisa dan mudah untuk dilakukan analisis dan di kategorikan sesuai dengan kata yang dibutuhkan Tabel 4.

Tabel 4. Penghapusan Tanda Baca

Awal	Akhir
rt indonesiagaruda: kini para penumpang economy class dari jakarta ke jepang, korea, australia & amsterdam juga sudah bisa #bookyourmeal	rt indonesiagaruda kini para penumpang economy class dari jakarta ke jepang korea Australia amp amsterdam juga sudah bisa bookyourmeal
facebook:<post> comments penerbangan jakarta (cgk) ke amsterdam (ams) free wifi juga ya	facebook post comments penerbangan jakarta cgk ke amsterdam ams free wifi juga ya

Setalah penghapusan tanda baca *(punctuation)* berikutnya adalah menghapus angka, dalam sebuah kata keberadaan angka kurang berpengaruh untuk dilakukan analisis teks, oleh karena itu perlu dihilangkan agar dapat mempermudah langkah berikutnya dalam menganalisis kata proses menghilangkan angka ditunjukan pada
Tabel *5*

Tabel 5. Menghilangkan Angka

Awal	Akhir
rt vivanewscom erick thohir bakal bubarkan 5 anak usaha garuda indonesia vivanews	rt vivanewscom erick thohir bakal bubarkan anak usaha garuda indonesia vivanews
pertama kali terbang pake bombardier dengan egpyt air dari sharm el sheihk ke kairo dulu tahun 2009 dan dengan garuda indonesia dari makassar ke surabaya dan semarang ...bravo garuda semoga segera bisa terbang ke eropa dan usa	pertama kali terbang pake bombardier dengan egpyt air dari sharm el sheihk ke kairo dulu tahun dan dengan garuda indonesia dari makassar ke surabaya dan semarang ...bravo garuda semoga segera bisa terbang ke eropa dan usa

Langkah selanjutnya adalah proses *filtering* yakni membuang daftar kata-kata yang kurang penting untuk di analisis menggunakan *stopwords*. *Stopword / stoplist* adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan *bag-of-words*. Jumlah stopwords yang digunakan adalah sejumlah 758 kata. Detail *stop word* yang digunakan ditunjukan pada Tabel *6*.

Tabel 6. *Stop Word*

No	Kata
1	ada
2	adalah
3	adanya
4	adapun
5	agak
6	agaknya
7	agar
8	akan
9	akankah
10	akhir
…..	……
757	yakni
758	yang

Menggunakan *stop word* untuk menghilangkang kata yang tidak diperlukan juga masih kurang, perlu ditambahkan penghilangan secara manual, atau langkah menambahkan kata ke dalam file *stop word* dapat dilakukan. Pada penelitian kali ini ditambahkan beberapa kata yang kurang bermanfaat untuk menggali informasi teks sesuai dengan konteks yang dibahas. Hal ini dapat berbeda-beda dari setiap penelitian, disini contoh kata *offcut* tidak mengandung makna apapun, karena hanya merupakan bagian informasi pengambilan data dari sosial media facebook, pemilihan *stop word* ini dapat disesuaikan dengan konteks masing-masing kata yang akan di analisis. Data kata tambahan ditunjukan pada Tabel *7*.

Tabel 7. *Stop Word* Tambahan

No	Kata
1	yang
2	*facebook*
3	*fetched*
4	*post*
5	*comments*
6	*data*
7	*with*
8	*offcut*
9	*twitter*
10	*nofollow*
…..	……
96	*dlvr*
97	*android*

Setelah menerapkan stop word ke dalam analisis text, langkah selanjutnya adalah menghapus spasi yang tidak berguna, yakni terdapat spasi yang berlebih pada selah antara dua kata. Contoh penearapan penghapusan spasi berlebih ditunjukan pada Tabel 8.

Tabel 8. Penghapusan Spasi Berlebih

Awal	Akhir
rt indonesiagaruda: kini para penumpang economy class dari jakarta ke jepang, korea, australia & amsterdam juga sudah bisa #bookyourmeal	rt indonesiagaruda kini penumpang economy class jakarta jepang korea Australia amsterdam sudah bisa bookyourmeal

Langkah yang terakhir sebelum dilakukan analisis *text mining*, perlu dilakukan penghapusan *URL web* agar text yang akan di analisis lebih mudah. Detail penghapusan url ditunjukan pada Tabel 9.

Tabel 9. Penghapusan *URL*

Awal	Akhir
RT @vivanewscom: Erick Thohir Bakal Bubarkan 5 Anak Usaha Garuda Indonesia https://t.co/JXkYgTHyRH #vivanews	RT @vivanewscom: Erick Thohir Bakal Bubarkan 5 Anak Usaha Garuda Indonesia #vivanews

Analisis Text Mining

Tahapan terakhir adalah dengan melakukan analisis *text mining* untuk dicari informasi mengenai kata yang paling sering muncul dan divisualisasikan ke dalam bentuk *word cloud*. Untuk mencapai tahapan ini, perlu dilakukan filtering dan *cleaning* data yang sudah dilakukan sebelumnya. Tujuannya adalah untuk menghindari kata-kata yang kurang memiliki makna dan karakter yang dianggap di luar konteks pembahasan.

a. Term Document Matrix

Term Document Matrix digunakan untuk mencari jumlah kata yang paling sering diutarakan oleh pengguna sosial media beserta jumlah frekuensinya. Pada sosial media *Twitter* kata yang paling sering muncul adalah Indonesia dengan frekuensi sebesar 1572. Detail pada *Term Document Matrix* ditunjukan pada Tabel *10*.

Tabel 10. *Term Document Matrix Twitter*

Kata	Frekuensi
indonesia	1572
garuda	1429
pssi	429
timnas	264
bali	213
yennywahid	208
pesawat	184
official	169
indonesiagaruda	147
digeeembok	118
anak	114
flight	114
dharma	113
nepal	112
bumn	111
erick	109
aryasinulingga	107

happy	106
mahashivratri	105
raghuna	105

Pada sosial media *Facebook* kata yang paling sering muncul adalah Garuda dengan frekuensi sebesar 3621. Detail pada *Term Document Matrix* ditunjukan pada Tabel 11

Tabel 11. *Term Document Matrix Facebook*

Kata	**Frekuensi**
garuda	3621
indonesia	1922
tiket	806
user	752
penerbangan	666
jakarta	575
rute	516
pesawat	461
harga	449
promo	424
naik	362
semoga	344
terbang	342
flight	341
selamat	323
kapan	299
berapa	281
hari	276
bagasi	268
maskapai	252

b. Word Cloud

Word Cloud berfungsi dalam mevisualisasikan kumpulan kata dalam *Term Document Matrix* menjadi sebuah tampilan yang lebih menarik. Adapun hasil dari platform twitter ditunjukan pada Gambar 2.

Gambar 2. Word Cloud Twitter

Adapun hasil dari *platform Facebook* ditunjukan pada Gambar 3.

Gambar 3. *Word Cloud Facebook*

c. Kesimpulan *Term Document Matrix* dan *Word Cloud*

Di *platform Twitter* kata yang banyak muncul lebih ke arah personal seperti nama personal yenny wahid, aryasinulingga, Erick, ada juga akun *twitter* digeeembok. Ini menjadi wajar ketika memang ada informasi yang sedang cukup banyak diperbincangkan berkaitan dengan maskapai penerbangan di Indonesia.

Pada platform *Facebook* banyak kata muncul berupa kata yang biasa digunakan untuk sebuah perusahaan transportasi seperti pesawat, rute, bagasi, maskapai, promo, harga dan kata-kata yang masih memiliki tendensi bernilai positif.

5. KESIMPULAN DAN SARAN

Kesimpulan yang dapat ditarik pada penelitian ini adalah sebagai berikut :

a. Kata yang paling sering muncul pada *platform Facebook* adalah garuda dengan frekuensi sebesar 3621 kata, sedangkan pada *Twitter* kata Indonesia sebesar 1572 kata.
b. *Facebook* memiliki kata-kata yang lebih cenderung positif, sedangkan *Twitter* lebih cenderung ke arah *sentiment* negatif.
c. Penerapan *term document matrix* dan *word cloud* saling melengkapi dalam proses *text mining.*

Adapun saran dalam penelitian ini adalah sebagai berikut :

a. Perlu dilakukan *labeling sentiment* untuk pengembangan penelitian selanjutnya
b. Perlu diterapkan algoritma klasifikasi untuk dapat dilakukan analisis yang lebih mendalam dan lebih baik lagi.

DAFTAR PUSTAKA

[1] V. Kale, "Enterprise performance intelligence and decision patterns," *Enterp. Perform. Intell. Decis. Patterns*, pp. 1–262, 2017, doi: 10.4324/9781351228428.

[2] D. Saputra, B. Pratama, Y. Akbar, W. G.-C. O. SPOT, and undefined 2018, "Penerapan Text Mining Untuk Assingment Complaint Handling Customer Terhadap Divisi Terkait Menggunakan Metode Decission," *Jurnal.Stikomcki.Ac.Id*, vol. 11, no. 2, pp. 207–216, 2018.

[3] 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.

[4] R. Siringoringo, "Text Mining dan Klasterisasi Sentimen Pada Ulasan Produk Toko Online," *Tek. Inform. Univ. Prima Indones. Medan*, vol. 2, pp. 1–6, 2019.

[5] M. Yasid, "Analisis Sentimen Maskapai Citilink Pada Twitter Dengan Metode Naãve Bayes," *J. Ilm. Inform.*, vol. 7, no. 02, p. 82, 2019, doi: 10.33884/jif.v7i02.1329.

[6] M. P. Simatupang and D. P. Utomo, "Analisa Testimonial Dengan Menggunakan Algoritma Text Mining Dan Term Frequency- Inverse Document Frequence (Tf-Idf) Pada Toko Allmeeart," *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 3, no. 1, pp. 808–814, 2019, doi: 10.30865/komik.v3i1.1697.

[7] T. S. Kartikasari *et al.*, "IMPLEMENTASI TEXT MINING UNTUK ANALISIS OPINI PUBLIK TERHADAP CALON PRESIDEN," vol. 7, no. 1, 2018.

[8] N. L. Ratniasih, M. Sudarma, and N. Gunantara, "Penerapan Text Mining Dalam Spam Filtering Untuk Aplikasi Chat," *Maj. Ilm. Teknol. Elektro*, vol. 16, no. 3, p. 13, 2017, doi: 10.24843/mite.2017.v16i03p03.

[9] F. Abdurrahman, M. Irfan, R. Andrian, J. A. H. N. No, and K. Bandung, "Implementasi Algoritma TF-IDF untuk Pencarian Pedoman Akademik dan Penentuan Sanksi Pada Jurusan Teknik Informatika UIN Sunan Gunung Djati Bandung," *Insight*, vol. 1, no. 1, pp. 133–140, 2017.

BIODATA PENULIS

Musthofa Galih Pradana, M.Kom.
Dosen Program Studi Informatika Universitas Alma Ata. Memperoleh Gelar Magister Komputer (M.Kom.) tahun 2019 dan Gelar Sarjana Komputer (S.Kom.) tahun 2017 di Universitas AMIKOM Yogyakarta.
Email : mgalihprada@uaa.ac.id