Jurnal IPTEK-KOM (Jurnal Ilmu Pengetahuan dan Teknologi Komunikasi) eISSN 2527 - 4902
Vol. 23 No. 2, Desember 2021: 141 - 153

Text Mining - Analisis Teks Terkait Isu Vaksinasi COVID-19

Text Mining - Text Analysis Related to COVID-19 Vaccination Issues

Novita Anggraini[1], Edi Surya Negara Harahap [2], Tri Basuki Kurniawan [3]
[1,2,3]Universitas Bina Darma, Jl. Jenderal A. Yani No. 3 Palembang Sumatera Selatan, Indonesia
Telp. 0711-515582

[1]novitaanggraini.opi@gmail.com, [2]edisuryanegaraharahapi@gmail.com, [3]tribasukikurniawan@binadarma.ac.id

Naskah diterima: 8 Agustus 2021, direvisi: 13 Desember 2021, disetujui: 14 Desember 2021

Abstract

As a step to reduce the transmission of COVID-19, the government is promoting a vaccination program to achieve herd immunity. Due to the failure of previous vaccinations, most of the people vehemently rejected vaccination, this was very unfortunate because there was a commotion in the community. In the process of regaining public trust, the government tries to disseminate vaccination information through social media (instagram), then this is what attracts researchers to further explore the vaccination process. From the many public opinions, there are some things that may still be difficult to find, because that is the need for text analysis. Text analysis was conducted to see the ranking terms and other information using the Rule-based Sentiment Analysis method. TF-IDF & LSI/LSA are types of rule mining methods used in the application of information extraction. The results of the analysis of this study may influence other information. For example, user perception analysis is used to see a broader picture of important issues or topics of conversation, as well as meeting points for problems related to COVID-19 vaccination.

Keywords: *COVID-19, Vaccination, Text Mining, TF-IDF, LSA, Sentiment Analysis.*

Abstrak

Sebagai langkah untuk dapat mengurangi penularan COVID-19, pemerintah tengah menggalakkan program vaksinasi sehingga tercapainya herd immunity. Disebabkan kegagalan vaksinasi sebelumnya, sebagian besar masyarakat menolak dengan keras adanya vaksinasi, hal ini sangat disayangkan karena terjadi kegaduhan ditengah–tengah masyarakat. Dalam proses menarik kembali kepercayaan masyarakat, pemerintah mecoba menyebaran luaskan informasi vaksinasi lewat media sosial (instagram), kemudian inilah yang menjadi daya tarik peneliti untuk mengeksplorasi lebih lanjut proses vaksinasi. Dari banyaknya opini masyarakat terdapat beberapa hal yang mungkin masih sulit ditemukan, sebab itulah perlunya analisis teks. Analisis teks dilakukan bertujuan melihat term rangking dan informasi lainnya dengan metode Rule-based Sentiment Analysis. TF-IDF & LSI/LSA adalah jenis metode rule mining yang digunakan dalam penerapan ektrasi informasi. Hasil analisis penelitian ini kemungkinan mempengaruhi informasi lainnya. Seperti analisis persepsi pengguna digunakan untuk melihat gambaran lebih luas tentang isu atau topik pembicaraan penting, serta titik temu permasalahan berkaitan dengan vaksinasi COVID-19.

Kata Kunci*:* *COVID-19,* Vaksinasi*, Text Mining, TF-IDF, LSA,* Analisis Sentimen*.*

PENDAHULUAN

Sosial media kini menjadi tempat perbincangan publik (Anggraini and Suroyo 2019). Hal ini sangat unik untuk diteliti, karena sebagian besar mengandung opini sentimen. Vaksinasi COVID-19 yang merupakan percobaan penanggulangan virus yang menyebar luas sejak akhir 2019 menjadi topik viral di jagad maya. Vaksinasi COVID-19 menjadi langkah yang digalakkan pemerintah untuk tercapainya *herd immunity*. Disebabkan kegagalan vaksinasi sebelumnya, sebagian besar masyarakat menolak dengan keras adanya vaksinasi. Hal ini sangat disayangkan karena terjadi kegaduhan ditengah–tengah masyarakat. Dalam proses menarik kembali kepercayaan masyarakat, pemerintah mecoba menyebaraluaskan informasi vaksinasi lewat sosial media salah satunya Instagram.

Analisis teks merupakan upaya untuk meninjau opini masyarakat terhadap suatu isu yang berkembang di media sosial. Analisis teks dilakukan untuk menghasilkan informasi spesifik (Negara, Andryani, and Saksono 2016) seperti melihat *term rangking* dan *isu* yang berkembang terkait proses vaksinasi serta informasi lainnya. Salah satu metode yang digunakan untuk melihat perkembangan isu adalah *Rule-based Sentiment Analysis*.

TF-IDF (*Term Frequency-Inverse Document Frequency*) & LSI/LSA *(Latent Semantic Indek/Analysis)* adalah jenis metode *rule mining* yang digunakan dalam penerapan ektraksi informasi. TF-IDF merupakan statistik numerik yang menunjukkan relevansi kata kunci dengan beberapa dokumen tertentu atau dapat dikatakan, menyediakan kata kunci tersebut, yang dengannya beberapa dokumen tertentu dapat diidentifikasikan atau dikategorikan (Qaiser and Ali 2018). TF-IDF tidak memuat jumlah kemunculan kata saja tetapi juga melihat kata-kata penting dan kurang penting dari dokumen yang berkaitan (Imamah and Hastarita Rachman 2020). TF-IDF juga digunakan untuk pencarian dan pengambilan informasi dari dokumen (Ramli et al. 2020). Kegunaan dasar TF_IDF adalah untuk menggambarkan pentingnya suatu *term* dalam sebuah dokumen dari korpus, yang kemudian digunakan sebagai istilah pembobotan dalam pencarian informasi, *text mining*, dan sebagainya. Dalam penelitian ini, TD_IDF digunkan untuk mencari temuan informasi berkaitan dengan isu, topik, dalam pembahasan proses vaksinasi COVID-19.

Penelitian ini berfokus pada pembahasan isu vaksinasi COVID-19. Jenis data komentar akan dilabeli apakah berupa kalimat itu positif, negatif, atau netral. Pelabelan bekerja dengan cara memahami makna kalimat berdasarkan konteks yang dibicarakan. Sehingga, kalimat dapat diidentifikasi masuk dalam kategori kelas sentiment yang seperti apa. Dengan *text mining*, kalimat akan dianalisis sentimennya. Sentimen analisis mempermudah dan mempercepat pencarian informasi yang dibutuhkan yang dilakukan, termasuk mengektraksi data dari sosial media (J. Samudra et al. 2009; Sutabri et al. 2018).

Penerapan analisis sentiment selalu berkaitan dengan pengkategorian kelas sentiman untuk menghasilkan informasi. Ada banyak metode yang dapat digunakan, salah satunya adalah *lexicon* atau *library* yang menilai kata demi kata. *Lexicon* sangat berguna tetapi hasil analisisnya tidak selalu bisa diterima dalam penerapan makna kalimat. Ketika kata digunakan dalam konteks kalimat yang berbeda maka akan berbeda pula makna kalimat yang dihasilkan. Oleh karena kelemahannya itu, penelitian ini menghindari penggunaan *lexicon* meski telah tersedia *tools* dari pihak ketiga. Algoritma *rule mining* dalam disiplin ilmu *text mining* lebih tepat untuk digunakan.

Selain TF-ID, salah satu produk sampingan dari LSI adalah daftar peringkat topik yang paling penting dalam sebuah *corpus* (Bader, Kegelmeyer, dan Chew 2011). Elemen pada matriks LSI mengandung frekuensi kemunculan kata pada setiap bagian dokumen (T.K. Landauer et al; Fernando and Toba 2020). Analisis teks membutuhkan *prepocess text* untuk menghasilkan data

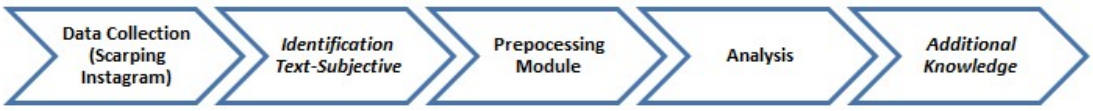
eISSN 2527 - 4902

bersih yang selanjutnya digunakan untuk mencari kata kunci (*key phrases*) di dalam dokumen. Oleh karena dokumen menggunakan bahasa Indonesia, maka *prepocess text* memanfaatkan *library sastrawi* yang menyediakan kata dalam bahasa (Andylibrian 2021). Proses ini menghasilkan data *cleaning* yang selanjutnya akan digunakan untuk mencari *key phrases* atau kata kunci dari sebuah dokumen, proses menggunakan model prediksi kata berikutnya atau urutan n item yang bedekatan yang disebut *n-gram*.

Langkah untuk mencari kata kunci dilakukan dengan 3 *n-gram* atau disebut *Trigram.* Trigram akan menentukan tiga item yang mungkin muncul berdekatan. *n-gram* menangkap konteks di mana kata-kata digunakan bersama (Subramanian 2019). Misalnya, mungkin ide yang baik untuk mempertimbangkan *Trigram* seperti "Sekolah Menengah Atas" atau "Kartu Tanda Penduduk" daripada memecahnya menjadi kata-kata individual seperti "Sekolah", "Menengah" dan "Atas". Berkaitan dengan itu *bag of word* digunakan untuk membantu kita mengubah kalimat teks menjadi vektor numerik. Model *Bag of Words* (BoW) adalah bentuk paling sederhana dari representasi teks dalam angka. Seperti dalam penggunaan TF-IDF yang mencari frekuensi *term*, *Bag of word* digunakan untuk membantu proses tersebut. Selain analisis persepsi penelitian ini mencoba melakukan keterkaitan informasi penting dengan bantuan *text mining* untuk mendapatkan informasi terkait proses vaksinasi *COVID-19.*

METODE

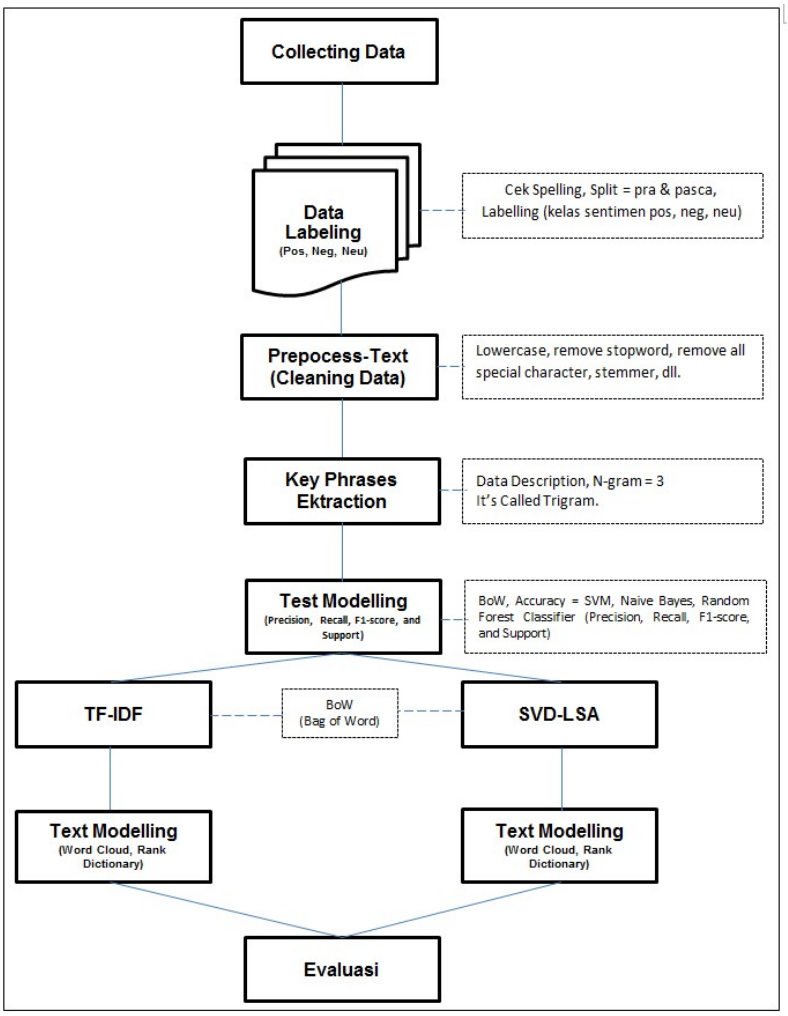
Penelitian ini menggunakan *metode Rule-based Sentiment Analysis* atau Analisis sentimen berbasis-aturan. Metode ini menggunakan algoritma *rule mining* guna menemukan fitur dari suatu produk lalu menemukan opini yang terkait dengan produk tertentu (Karim and Das 2018; Anggraini and Suroyo 2019). Untuk melakukan *Rule-based Sentiment Analysis*, dapat mengikuti prosedur berikut (Kundi et al. 2014; Anggraini and Suroyo 2019) yang disesuaikan dengan alur kerja penelitian ini:



Gambar 1. Prosedure Kerja *Rule-Based Sentimen Analysis*
(Kundi et al. 2014; Anggraini and Suroyo 2019)

Teks yang digunakan adalah komentar pengguna, *posting*, dan data lain yang ada di akun Instagram @kemenkes_ri (*https://www.Instagram.com/kemenkes_ri*). Data diambil beberapa bulan sejak waktu publikasi vaksinasi sendiri. Format data adalah *csv. Data komentar diambil setiap bulannya dan digabungkan mnejadi satu file dengan perintah *command prompt.* Sedangkan data lainnya akan dipisah sesuai analisis yang akan dilakukan. Penelitian ini berfokus pada pembahasan isu dalam perspektif pengguna, bukan pembahasan mengenai proses pengkategorian perspektif mana positif dan mana negatif dan bukan pula mengenai berapa nilai yang diberikan pada setiap kata karena hal tersebut telah dilakukan dalam proses pelabelan itu sendiri.

Tahapan analisis teks dalam penelitian ini dapata terlihat di Gambar 3. Penelitian menerapkan 5 (lima) tahapan yaitu *data labeling, preprocess-text, key phrases, test modeling,* dan *tex modeling.*



Gambar 3. Instrumen Penelitian

Data labeling adalah tahap memberikan label pada data yaitu komentar di akun @kemenkes_ri. Jenis data komentar yang diambil akan diubah menjadi *dataset* yang kemudian akan dilakukan pelabelan pada setiap baris komentar data tersebut apakah kalimat itu positif, negatif atau netral. Cara kerja pelabelan sendiri dengan memahami makna kalimat berdasarkan konteks yang dibicarakan bukan penilaian kata per kata. Tabel 1 menunjukkan panduan proses pelabelan. Dari *data labeling* didapat dataset yang akan diolah di tahap selanjutnya. Dataset *data labeling* dapat dilihat di Tabel 2.

Tabel 1. Panduan Kerja Proses Pelabelan

No	Kelas Sentimen	Kondisi
1.	Positif	Dukungan, saran membangun, cinta, doa, pertanyaan jadwal vaksin, kuota vaksin
2.	Negatif	Ketidakpercayaan, pernyataan penolakkan, fitnah, hoax, dan cacat dari proses vaksinasi itu sendiri, bisa berupa sindiran, dan peralihan keputusan vaksin atau saran vaksin untuk orang dengan alasan nanti melihat proses vaksinasi yang dilakukan apakah, baik atau buruk, aman atau tidak, barulah dia mengambil keputusan dan keluhan dari proses vaksinasi itu sendiri.
3.	Netral	Komentar penyisihan, mereka yang memihak atau tidak memihak proses vaksinasi, seputar pertanyaan yang memang menginginkan info atau jawaban untuk membuat keputusan, tagging, iklan, cerita pasien diluar vaksinasi, sapaan.

Tabel 2. Deskripsi Dataset

Kriteria	Dataset (komentar)
Columns	mediaUrl, username, userUrl, text, class
Used Columns	text, class
Rows	1,532 rows
Rows pasca-cleaning	1,374 rows
Positif	592 rows
Negatif	407 rows
Netral	375 rows
Token/Key phrases	19077 item

Tahap *prepocess-text* bertujuan untuk mengolah dataset agar dapat dibaca oelh algoritma dan mengektraksi informasi yang di inginkan. *Preprocess-text* akan menghasilkan data bersih atau *clean-dataset* yang berisi informasi spesifik. Prosedur di tahap ini meliputi:

1. Menghapus semua karakter khusus seperti *emoticon* "☺" dan "☹". Komentar yang hanya mengandung *emoticon* juga dihapus karena dalam kasus *term issue* hal ini kurang begitu penting untuk dianalisis.
2. Mengganti spasi ganda dengan spasi tunggal.
3. Mengubah semua huruf menjadi huruf kecil (*lowercase*).
4. Menghapus *stopword*. *Stopword* atau kata yang sering muncul sering kali menjadi titik masalah, maka dalam penerapan *prepocess text* diberi kondisi untuk menentukan mana yang perlu masuk kedalam *stopword*. Contohny adalah "ada", "adalah", "adanya", "adapun", "apakah", "apalagi", "apatah", "artinya", "asal", "asalkan", "atas", "atau", "ataukah", "ataupun", "awal", "awalnya", "bagai", "yang", "dan", "atau", dan "sebagainya".
5. Mengembalikan kata dalam bentuk kata dasar (*stemmer*).

Selanjutnya adalah tahap *key pharases extraction* yaitu penggunaan *n-gram* untuk mendapatkan *key pharses* atau kata kunci. Data dibuat menjadi token-token atau kata kunci (*key pharses*) yang selanjutnya akan dimasukkan ke dalam *bag of word* untuk dikompres menjadi bentuk vektor dalam proses selanjutnya (*text testing* dan *text modelling*). Penggunaan *trigram* bertujuan tidak menghilangkan informasi tersembunyi. Sebagai contoh "Sekolah Menengah Atas" dan "Kartu Tanda Penduduk" jika penggunaan hanya *onegram* atau *bigram* maka kasus diatas tidak terdeteksi. Atau contoh lain adalah "membaca koran buku" jika dikaitan dengan kedekatan item/kata maka ditampilkan kata kunci yang serupa dalam dokumen tersebut. Membaca buku, membaca koran, buku dan koran sama-sama benda yang untuk dibaca.

Kata kunci yang yang dihasilkan kemudian dikompres dalam bentuk vektor di tahap *test modelling.* Kata kunci yang masuk ke dalam *bag of word* akan dipetakan dan dikompres dalam bentuk vector agar mudcah dibaca. Untuk mendapatkan akurasi data, *test modelling* menggunakan teknik *Support Vector Machine*, *Naive Bayes,* dan *Random Forest Clasifier*. Perbandingan akurasi ketiga teknik tersebut dapat terlihat di Tabel 3.

Tabel 3. Accuracy dengan SVM, Naive Bayes, RFC

	Support Vektor Machine (SVM)	Naive Bayes	Random Forest Clasifier (RFC)
Dataset	0.46	0.85	0.96

Di tahap *test modelling* juga dilakukan TF-IDF dan SVD. Tiga aspek yang mempengaruhi skema pembobotan *term* adalah *term frequency* (TF), *inverse document frequency* (IDF), dan normalisasi. Dalam penelitian ini, TF-IDF digunakan sebagai metode pembobotan fitur yang menghasilkan informasi. TF-IDF menentukan bobot *term* dengan dua faktor yaitu:

1. Menghitung jumlah kemunculan *term* j pada dokumen i atau disebut juga dengan frekuensi *term* yang dilambangkan dengan tf_{ij}.
2. Menghitung frekuensi kemunculan pada seluruh kumpulan dokumen atau disebut juga dengan frekuensi dokumen yang dilambangkan dengan df_{ij}.

Pembobotan term TF-IDF terbentuk pada rumus berikut, yaitu:

$$TF.IDF = tf_{ij} x IDF_{ij} = TF_{ij} x log \frac{N}{DF_j} \quad \text{.....(1)}$$

Di mana:
N = jumlah dokumen dalam koleksi
TF = *term frequency*
IDF = *inverse document frequency*.

Term Frequency (TF) digunakan untuk mengukur berapa kali suatu *term* hadir dalam suatu dokumen (Hakim, A. A. et al. 2015; Qaiser dan Ali 2018). Sebagai permisalan, kita memiliki dokumen "NOV" yang berisi 10,000 kata dan kata "Vaksin" hadir dalam dokumen tepat 50 kali. Fakta yang sangat diketahui bahwa, untuk panjang total dokumen bisa bervariasi dari sangat kecil hingga besar, karena itulah ada kemungkinan istilah apa pun dapat muncul lebih sering dalam dokumen besar dibandingkan dengan dokumen yang lebih kecil. Maka dari itu, untuk memperbaiki masalah ini, kemunculan istilah apa pun dalam dokumen dibagi dengan total istilah yang ada dalam dokumen itu, untuk menemukan frekuensi istilah. Sehingga, dalam hal ini frekuensi istilah kata "Vaksin" dalam dokumen "NOV" adalah;

TF(w) =
(berapa kali kata w muncul dalam dokumen) / (jumlah total kata dalam dokumen)(2)
TF = 50/10,000 = 0,005

Ketika frekuensi istilah dokumen dihitung, dapat diamati bahwa algoritma memperlakukan semua *keyword* secara setara, tidak masalah jika itu adalah *stopword* seperti "dan". Semua kata kunci memiliki kepentingan yang berbeda. Katakanlah, *stopword* "dan" hadir dalam dokumen 5000 kali tetapi tidak ada gunanya atau memiliki arti yang sangat kurang, itulah gunanya IDF. Frekuensi dokumen terbalik memberikan bobot yang lebih rendah untuk kata-kata yang sering muncul dan memberikan bobot yang lebih tinggi untuk kata-kata yang jarang muncul. Misalnya, kita memiliki 16 dokumen dan istilah “obat” hadir dalam 8 dokumen tersebut sehingga frekuensi dokumen terbalik dapat dihitung sebagai;

IDF(w) = log(jumlah total dokumen / jumlah dokumen dengan w di dalamnya)..... (3)
IDF = log_e (16/8) = 0,3010

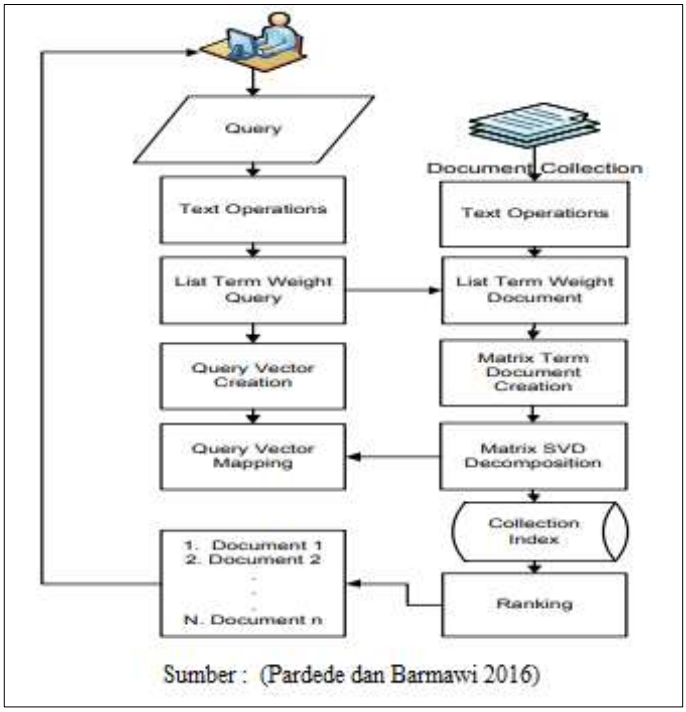
Semakin besar atau tinggi kemunculan dari kata dalam dokumen akan memberikan frekuensi istilah yang lebih tinggi (TF) dan semakin sedikit kemunculan dari kata dalam dokumen akan menghasilkan tingkat kepentingan yang lebih rendah (IDF) untuk kata kunci yang dicari dalam dokumen tertentu. Maka untuk menghitung TF-IDF dilakukan dengan formula:

TF-IDF(w) = TF(w) x IDF(w) .....(4)
TF-IDF = 0.005*0.3010 = 0.001505

LSI/LSA adalah algortima kedua yang dipakai untuk analisis teks. Metode LSI/LSA merupakan metode yang diimplementasikan dalam Sistem IR untuk mencari dan menemukan informasi berdasarkan keseluruhan makna dokumen bukan hanya makna individu kata-kata. LSI/LSA juga merupakan metode yang mengidentifikasi hubungan kontekstual antar kata dalam

eISSN 2527 - 4902

sebuah kalimat (Rahmalia 2020). Dalam metode LSI, pengumpulan dokumen dibangun dalam bentuk ruang vektor dengan menggunakan bantuan *Singular Value Decompositions* (SVD) (Rosario, B 2000; Pardede and Barmawi 2016). Alur proses dalam metode LSI dapat jabarkan pada Gambar 2.



Gambar 2. Alur Proses LSI (Pardede and Barmawi 2016).

Cara kerja LSA/LSI adalah dengan menghasilkan model yang didapatkan dengan mencatat kemunculan-kemunculan kata dari tiap-tiap dokumen yang direpresentasikan ke dalam sebuah matriks yang dinamakan *term-document matrix*. Proses *Singular Value Decomposition* (SVD) digunakan untuk mendapatkan *Cosine Similarity* (nilai kemiripan) antara satu dokumen dengan dokumen yang lainnya (Landauer and Dumais, 1997; Landauer, Foltz, and Laham, 1998; Wonowidjojo and Hartono 2016).

SVD merupakan cara dekomposisi matriks yang digunakan untuk mencari kesamaan antar segmen kata**.** SVD merupakan komponen pemrosesan yang mengkompresi informasi yang berkaitan dalam jumlah besar ke dalam ruang yang lebih kecil (Dary 2015). Formula dari SVD adalah sebagai berikut:

$$A = USV^T \quad \text{..... (5)}$$

Di mana :
A : matriks asal
U : *orthonormal eigenvector* dari AAT
S : matriks diagonal
VT : transpose dari *orthogonal* matriks V

Cosine similarity adalah persamaan kosinus antara vektor dokumen dan document vektor queri dihitung sebagai berikut:

$$similarity\ (Q, D_j) = \cos(Q, D_j) = \frac{Q.D_j}{|Q||D_j|} = \frac{1}{|Q||D_j|} = \sum_{i=1}^{r} q_j * d_{ji} \quad \text{.....(6)}$$

Proses akurasi dilakukan dengan SVM, Naive Bayes, Random Forest Clasifier. Yang meliputi P*recision, Recall, F1-Score, Support.*

a. *Precision*
   Presisi = positif benar / jumlah prediksi. .....(7)
b. *Recall*
   Recall = true positive / actual count......(8)
c. F1-*score*
   *F1-score = 2 * precision * recall / (precision + recall).* .....(9)
d. *Support*
   Jumlah kemunculan setiap kelas di y_test (*matrix code*). .....(10)

Setelah diperoleh *bag of word*, data kemudian diranking di tahap *text modelling.* Data yang telah diberi label positif, negatif, dan netral akan dinilai dalam kategori kelas yang bersangkutan. Sehingga dapat dilihat kata kunci mana yang mewakili isu penting yang dibicarakan oleh pengguna dalam kumpulan dokumen komentar tersebut.

HASIL DAN PEMBAHASAN

Setelah dilakukan analisis dengan TF-IDF dan SVD-LSA, didapat kata kunci yang merupakan titik tumpu penemuan permasalahan seperti terlihat di Tabel 4 dan Tabel 5. Dari hasil analisis yang dilakukan diketahui bahwa TF-IDF menemukan kata kunci yang sering disinggung komentator di akun @kemenkes_ri karena sering muncul. Akan tetapi, kita tidak bisa mengetahui detail pembicaraan karena banyaknya dokumen data. Inilah masalah jika mengacu pada kata kunci yang dihasilkan. Kata kunci yang banyak muncul adalah “vaksin” dan “covid”. Kata ini muncul di semua isu baik positif, negatif, maupun netral. TF-IDF tidak memungkinkan kita untuk menemukan detail masalah karena banyaknya data. Semakin banyak data, semakin sulit ditemukan detail masalahnya. Namun, jika untuk melihat ranking isu, TF-IDF sangat membantu.

Tabel 4. Text Modelling TF-IDF

No	Issues # positif	Issues # negative	Issues # netral
1	('vaksin', 0.6910)	('vaksin', 0.7737)	('vaksin', 0.6858)
2	('semoga', 0.3004)	('covid', 0.2463)	('min', 0.2732)
3	('covid', 0.2628)	('orang', 0.1542)	('covid', 0.2620)
4	('virus', 0.1389)	('rakyat', 0.1194)	('orang', 0.1226)
5	('indonesia', 0.1314)	('negara', 0.0970)	('kemenkes_ri', 0.1170)
6	('kesehatan', 0.1164)	('sehat', 0.0970)	('aman', 0.0947)
7	('masyarakat', 0.0976)	('indonesia', 0.0870)	('mohon', 0.0947)
8	('orang', 0.0976)	('presiden', 0.0870)	('divaksin', 0.0892)
9	('min', 0.0938)	('masyarakat', 0.0845)	('efek', 0.0836)
10	('sehat', 0.0938)	('pakai', 0.0845)	('halal', 0.0836)
11	('masker', 0.0901)	('tuan', 0.0821)	('pasien', 0.0836)
12	('pakai', 0.0826)	('divaksin', 0.0796)	('gimana', 0.0724)
13	('pandemi', 0.0826)	('sakit', 0.0721)	('indonesia', 0.0724)

14	('imun', 0.0788)	('pemerintah', 0.0696)	('positif', 0.0724)
15	('pemerintah', 0.0788)	('kesehatan', 0.0671)	('vaksinasi', 0.0724)
16	('cepat', 0.0713)	('menteri', 0.0671)	('corona', 0.0669)
17	('corona', 0.0713)	('nakes', 0.0671)	('kesehatan', 0.0669)
18	('vaksinasi', 0.0713)	('dpr', 0.0646)	('negara', 0.0669)
19	('divaksin', 0.0676)	('pejabat', 0.0597)	('pakai', 0.0669)
20	('gratis', 0.0676)	('corona', 0.0572)	('antigen', 0.0613)
21	('semangat', 0.0676)	('nya', 0.0572)	('bayar', 0.0613)
22	('normal', 0.0638)	('duluan', 0.0547)	('anak', 0.0557)
23	('kemenkes_ri', 0.0600)	('korupsi', 0.0547)	('imunisasi', 0.0557)
24	('tuan', 0.0600)	('masker', 0.0547)	('nya', 0.0557)
25	('masuk', 0.0563)	('no', 0.0547)	('penyakit', 0.0557)
26	('mudah', 0.0563)	('beli', 0.0522)	('sakit', 0.0557)
27	('vaksin covid', 0.0450)	('vaksin covid', 0.0373)	('et al', 0.0501)
28	('pakai masker', 0.0413)	('vaksin vaksin', 0.0373)	('vaksin covid', 0.0501)
29	('terima kasih', 0.0375)	('pakai masker', 0.0348)	('efek samping', 0.0334)
30	('hidup normal', 0.0300)	('wakil rakyat', 0.0298)	('terima kasih', 0.0334)
31	('jaga jarak', 0.0262)	('beli vaksin', 0.0248)	('rapid antigen', 0.0278)
32	('uji klinis', 0.0262)	('anggota dpr', 0.0223)	('alat bantu', 0.0223)
33	('vaksin vaksin', 0.0262)	('jaga jarak', 0.0223)	('alat bantu dengar', 0.0223)
34	('virus covid', 0.0262)	('rumah sakit', 0.0223)	('bantu dengar', 0.0223)
35	('vaksin gratis', 0.0225)	('presiden menteri', 0.0199)	('free bayar', 0.0223)
36	('efek samping', 0.0187)	('uji coba', 0.0174)	('log unit', 0.0223)
37	('izin repost', 0.0187)	('vaksin sinovac', 0.0174)	('min vaksin', 0.0223)
38	('masyarakat vaksin', 0.0187)	('ahli teknologi', 0.0149)	('pakai vaksin', 0.0223)
39	('pasien positif', 0.0187)	('ahli teknologi laboratorium', 0.0149)	('rumah sakit', 0.0223)
40	('tenaga medis', 0.0187)	('butuh vaksin', 0.0149)	('vaksin aman', 0.0223)
41	('antigen virus', 0.0150)	('efek samping', 0.0149)	('vaksin halal', 0.0223)
42	('lepas masker', 0.0150)	('laboratorium medis', 0.0149)	('alergi obat', 0.0167)
43	('orang orang', 0.0150)	('laboratorium medis atlm', 0.0149)	('buah apel', 0.0167)
44	('pola hidup', 0.0150)	('medis atlm', 0.0149)	('covid min', 0.0167)
45	('positif covid', 0.0150)	('orang sehat', 0.0149)	('log unit reduction', 0.0167)
46	('suntik vaksin', 0.0150)	('sehat vaksin', 0.0149)	('mohon info', 0.016)
47	('vaksin lepas', 0.0150)	('teknologi laboratorium', 0.0149)	('mohon maaf', 0.0167)
48	('vaksin lepas masker', 0.0150)	('teknologi laboratorium medis', 0.0149)	('negara vaksin', 0.0167)
49	('akibat corona', 0.0112)	('tolak vaksin', 0.0149)	('of incubation', 0.0167)
50	('amin yrb', 0.0112)	('uji klinis', 0.01492)	('positif covid', 0.0167)

Tabel 5. SVD-LSA Text Modelling

No	Issues # positif	Issues # negative	Issues # netral
1	('min masyarakat', 0.0302)	('gratis', 0.0266)	('ditunggu', 0.0279)
2	('diterima', 0.0295)	('gimana', 0.0264)	('dites', 0.0277)
3	('covid bisnis', 0.0293)	('pejabat', 0.0258)	('pokoknya', 0.0267)
4	('menggiurkan', 0.0286)	('vaksin covid', 0.0256)	('mencegah', 0.0248)
5	('lanjutkan', 0.0282)	('no', 0.0253)	('oke', 0.0244)
6	('mendaftar', 0.0275)	('duluan', 0.0252)	('makasih', 0.0242)
7	('membicarakan', 0.0267)	('bayar', 0.0252)	('peduli', 0.0241)
8	('laksanakan', 0.0267)	('obat', 0.0252)	('dpr min', 0.0235)
9	('kepastian', 0.0267)	('pandemi', 0.0249)	('sehat vaksin', 0.0231)
10	('covid sih', 0.0264)	('cepat', 0.0249)	('merata', 0.0230)
11	('harganya', 0.0259)	('imunisasi', 0.0248)	('pemimpin', 0.0229)
12	('wajib sih', 0.0252)	('sih', 0.0248)	('wabah', 0.0228)
13	('vaksin wajib', 0.0250)	('coba', 0.0245)	('negara pakai', 0.0225)
14	('gratis bayar', 0.0242)	('semangat', 0.0245)	('dijamin', 0.0224)
15	('dpr rakyat', 0.0227)	('butuh', 0.0243)	('vaksin min', 0.0223)

16	('cepat pandemi', 0.0225)	('suntik', 0.0242)	('diutamakan', 0.0220)
17	('dpr menteri', 0.0224)	('penyakit', 0.0240)	('pertanyaannya', 0.0216)
18	('gratisnya', 0.0222)	('percaya', 0.0239)	('teridentifikasi', 0.0216)
19	('corona vaksin', 0.0221)	('kemenkes_ri', 0.0239)	('kebijakan', 0.0215)
20	('dibelinya', 0.0214)	('pakai', 0.0238)	('imunisasinya', 0.0214)
21	('emang wajib', 0.02146)	('vaksinasi', 0.0238)	('lupakan', 0.0214)
22	('datangnya', 0.0213)	('menteri', 0.0236)	('menolak', 0.0206)
23	('dipenjara', 0.0213)	('vaksinnya', 0.0236)	('hamil', 0.0206)
24	('aman nih', 0.0212)	('maaf', 0.0236)	('lucu', 0.0204)
25	('rbnaaaashshiddiqie', 0.0209)	('silahkan', 0.0236)	('infonya', 0.0204)
26	('menteri kakak', 0.0208)	('corona', 0.0236)	('meremehkan', 0.0203)
27	('vaksinasi min', 0.0207)	('nya', 0.0235)	('yra', 0.0202)
28	('nakes banget', 0.0206)	('vaksin vaksin', 0.0204)	('wajib', 0.0202)
29	('percaya corona', 0.0206)	('pakai vaksin', 0.0196)	('jualan', 0.0202)
30	('jurnalnya', 0.0205)	('beli vaksin', 0.0189)	('rakyat vaksin', 0.0201)
31	('nbilazzahr', 0.0203)	('vaksin gratis', 0.0181)	('berani', 0.0201)
32	('parnaidanatali_', 0.0203)	('vaksin aman', 0.0180)	('vaksin gratis', 0.0200)
33	('cksin', 0.0202)	('uji coba', 0.0175)	('vaksin bayar', 0.0200)
34	('jordifalah', 0.0202)	('tolak vaksin', 0.0174)	('pandemi cepat', 0.0196)
35	('pfitzer', 0.0202)	('pakai masker', 0.0174)	('salam sehat', 0.0193)
36	('apa^{o}', 0.0202)	('sehat vaksin', 0.0171)	('presiden menteri', 0.0192)
37	('indah_ekaaa', 0.0201)	('uji klinis', 0.0165)	('orang vaksin', 0.0192)
38	('menyimak', 0.0201)	('wakil rakyat', 0.0163)	('mirzaa no', 0.0188)
39	('aridarizk', 0.0201)	('vaksin nya', 0.0161)	('coba kali', 0.0186)
40	('allhamdulilah', 0.0201)	('presiden menteri', 0.0159)	('izin share', 0.0181)
41	('godblesstico', 0.0200)	('kalo vaksin', 0.0157)	('vaksin presiden', 0.0175)
42	('maaf tuan', 0.0200)	('vaksin halal', 0.0156)	('tuan nanya', 0.0172)
43	('butuh maaf', 0.0200)	('vaksin sinovac', 0.0156)	('vaksin lantas', 0.0170)
44	('cksinasi', 0.0200)	('kena covid', 0.0155)	('bayar aman', 0.0170)
45	('hamil aman', 0.0197)	('hidup normal', 0.0152)	('maaf bayar', 0.0170)
46	('aman gratis', 0.0196)	('virus covid', 0.0150)	('maaf bayar aman', 0.0170)
47	('pandemi allah', 0.0196)	('vaksin duluan', 0.0150)	('serius nanya', 0.0169)
48	('sehat sehat', 0.0194)	('vaksin negara', 0.0150)	('covid cepat', 0.0169)
49	('semangat anak', 0.0193)	('china china', 0.01500)	('gratis versi', 0.0168)
50	('presiden tuan', 0.0191)	('terima kasih', 0.0149)	('vaksin gratis versi', 0.0168)

Hasil temuan dengan metode LSI/LSA cukup menarik dibandingkan denga TF-IDF. LSI/LSA membantu menemukan detail pembicaraan dengan cepat. Kata kunci menunjukkan isu yang paling banyak dibicarakan atau makna tersembunyi dokumen tersebut bukan mengacu ke kata terbanyak. Hanya saja, perwakilan kata kuncinya cukup unik karena LSA memberikan kata kunci untuk menemukan informasi berdasarkan keseluruhan makna dokumen bukan hanya makna individu kata-kata. Misal TF-IDF sering membahas “vaksin” di rangking satu tetapi LSA memberi kata kunci “gratis” dalam kategori kelas negatif. Jika membahas tentang kelas negatif apa yang sering dibicarakan/diperdebatkan oleh komentator di akun @kemenkes_ri? Jika TF-IDF memberikan jawaban “vaksin” tentu saja semua dokumen rata-rata membahas “vaksin”. Namun, informasi tersembunyi dari dokumen tersebut apa selain ”vaksin” itu sendiri, tidak dapat diketahui.

Ketika menilik kata satu persatu, muncul pertanyaan selanjutnya. Lalu, gambaran permasalahan dari kata kunci “vaksin” itu sendiri apa? Jika dokumen terlalu banyak maka akan sulit untuk memahami dokumen data dan mengumpulkan informasi titik tumpu masalah berdasarkan kata kunci “vaksin” dalam kategori kelas sentimen negatif. Sedangkan LSA memberikan informasi tersembunyi dari kumpulan dokumen tersebut. Isu yang paling sering

dibahas apa? Kata-kata unik seperti "gratis" adalah salah satu kontribusi LSA. LSA dapat menemukan inti masalah dari kata "gratis" karena berada di kategori kelas sentimen negatif. Begitulah penjelasannya.

Penggabungan antara metode TF-IDF dan LSA mampu menemukan detail permasalahan. Kata "vaksin" di TF-IDF menempati rangking tearatas untuk semua isu, baik positif, negatif, maupun netral. Di LSA, kata "gratis" menempati rangking tertinggi di isu negatif. Maka, salah satu titik tumpu permasalahan yang sering dibahas adalah kenapa vaksin yang digunakan berbeda, mau vaksin gratis atau berbayar beberapa masyarakat tetap menolak, dan sebagainya. Selain itu jika ditelusuri ada perbedaan yang mencolok di antara keduanya, LSA memberikan nilai kemunculan kata kunci "vaksin" di bawah 0,03. Sedangkan nilai kemunculan di TF-IDF adalah 0,77. Artinya, 77 persen data membahas tentang "vaksin" itu sendiri tapi detail masalahnya tidak dapat diketahui melalui metode TF-IDF. Maka dari itu, LSA berguna untuk melihat lebih dalam lagi apa isu yang dibicarakan? Misalnya, ketika membicarakan "vaksin" pasti luas pembahasannya dan LSA memberikan detailnya. Bahwa ketika kata "vaksin" dibawa ke LSA akan mengarah ke beberapa kata kunci yang spesifik membahas "vaksin". Sebagai contoh LSA memberi jawaban tentang perbedaan jenis vaksin dan penolakkan masyarakat tentang program vaksinasi. Ini adalah kombinasi yang menarik dan baik untuk penelitian ini.

Dari analisis, diketahui bahwa isu yang sering dibicarakan adalah tentang vaksin COVID-19 dengan detail permasalahan diantaranya berbayar atau gratis, perbedaan jenis vaksin, efek samping, kandungannya, *trust issue* baik kepada vaksin, maupun pemerintahan, tuntutan pejabat sebagai garda terdepan, HAM, kebijakan pemerintah, penolakan, *hoax*, dan sebagainya. Salah satu yang membuat pengguna menolak vaksin adalah kurangnya pendidikan atau edukasi terhadap vaksinasi itu sendiri, seperti manfaat vaksin COVID-19, kandungannya, proses vaksinasi yang tidak jelas, dan *hoax*. Isu ini dapat menyebabkan kesalahpahaman. Terutama kondisi *dataset* di mana komentar negatif pada dokumen data lebih banyak mengarah ke praduga karena telatnya edukasi dan informasi yang diberikan atau bahkan masih termakan *hoax* dan lemahnya *trust issue* kepada pemerintahan (ini melihat dari data *posting* @kemenkes_ri). Jika dilihat rentang waktu publikasi dan edukasi yang disampaikan, terlihat ada kesenjangan antara informasi yang diterima dan informasi yang disebar. Beberapa orang menganggap vaksin itu berbayar namun ternyata informasi tentang vaksin gratis baru dipublikasi beberapa bulan kemudian. Atau, publikasi uji klinis yang sangat terlambat padahal vaksinasi telah dimulai. Hal ini membuat kepercayaan masyarakat kepada pemerintah berkurang. Belum lagi hal-hal lain seperti pengaruh sosial media dan pihak luar yang menjadi kambing hitam penyebar *hoax*.

KESIMPULAN

Penelitian ini berhasil menangkap isu penting dengan kombinasi metode TF-IDF dan LSA. Jika kedua metode ini digunakan bersama, maka akan diperoleh kata kunci yang menunjukan isu melalui TF-IDF dan makna tersembunyi atau detail permasalahan dari kata kunci melalui LSA. Dari analisis yang dilakukan diketahui bahwa *isu* yang sering dibicarakan adalah tentang vaksin COVID-19 dengan detail permasalahan adalah isu penolakan, perbedaan jenis vaksin, efek samping, kandungan, bisnis pemerintah, *trust issue*, berbayar atau gratis dan sebagainya. Salah satu yang membuat masyarakat menolak vaksin adalah kurangnya pendidikan atau edukasi terhadap vaksinasi, seperti manfaat vaksin COVID-19, kandungan vaksin, proses vaksinasi yang tidak jelas, dan *hoax*. Ada kesenjangan informasi antara pemerintah dan masyarakat yang menimbulkan

kesalahpahaman.

Untuk penelitian lebih lanjut, disarankan melalukan pembahasan pada data yang bersifat *uncomplete*. Agar dapat melakukan evaluasi *real-time,* maka penggunaan prediksi kelas sentimen dibutuhkan. LSA dapat digunakan untuk memprediksi kedekatan atau kemiripan. Hasilnya akan memberikan dampak yang luar biasa dalam kontribusi penelitian terutama tentang waktu. Dalam kasus ini LSA dapat diandalkan, selain itu ketika model tidak memiliki tingkat akurasi yang baik maka klasifikasi dengan LSA sangat membantu untuk mendongkrak akurasi dari model yang telah ada itu sendiri.

UCAPAN TERIMA KASIH

Terima kasih kepada Bapak Tri Basuki Kurniawan, S.Kom., M.Eng. Ph.D yang memberikan kontribusi arahan kepada jalannya penelitian ini. Terima kasih kepada Dr. Edi Surya Negara, M.Kom. selaku dosen pembimbing, yang membimbing saya hingga penelitian ini dapat terselesaikan dengan baik.

DAFTAR PUSTAKA

Andylibrian. 2021. Sastrawi. PHP. Sastrawi. https://github.com/sastrawi/sastrawi/blob/09db1bda7756fae740767ed7eb8de1b01ae859d5/README.en.md.

Anggraini, Novita, and Heri Suroyo. 2019. “Comparison of Sentiment Analysis against Digital Payment ‘T-cash and Go-pay’ in Social Media Using Orange Data Mining.” Journal of Information Systems and Informatics 1 (2): 152–63. https://doi.org/10.33557/journalisi.v1i2.21.

Bader, Brett W., W. Philip Kegelmeyer, and Peter A. Chew. 2011. “Multilingual Sentiment Analysis Using Latent Semantic Indexing and Machine Learning.” Dalam 2011 IEEE 11th International Conference on Data Mining Workshops, 45–52. Vancouver, BC, Canada: IEEE. https://doi.org/10.1109/ICDMW.2011.185.

Dary, Mochamad Irfan. 2015. “Analisis dan Implementasi Short Text Similarity dengan Metode Latent Semantic Analysis Untuk Mengetahui Kesamaan Ayat al-Quran,” 8.

Fernando, Edward Hanafi, and Hapnes Toba. 2020. “Pemanfaatan Latent Semantic Indexing untuk Mengukur Potensi Kerjasama Jurnal Ilmiah Lintas Universitas.” Jurnal Teknik Informatika dan Sistem Informasi 6 (3). https://doi.org/10.28932/jutisi.v6i3.2894.

Imamah, and Fika Hastarita Rachman. 2020. “Twitter Sentiment Analysis of COVID-19 Using Term Weighting TF-IDF And Logistic Regresion.” 2020 6th Information Technology International Seminar (ITIS), no. COVID-19, analysis sentiment (Oktober). https://doi.org/10.1109/ITIS50118.2020.9320958.

Negara, Edi Surya, Ria Andryani, and Prihambodo Hendro Saksono. 2016. “Analisis Data Twitter: Ekstraksi dan Analisis Data Geospasial.” Jurnal INKOM 10 (1): 27. https://doi.org/10.14203/j.inkom.433.

Pardede, Jasman, and Mira Musrini Barmawi. 2016. “Implementation of LSI Method on Information Retrieval for Text Document in Bahasa Indonesia.” INTERNETWORKING INDONESIA JOURNAL, 8 (1). ISSN 1942-9703.

Qaiser, Shahzad, and Ramsha Ali. 2018. "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents." International Journal of Computer Applications 181 (1): 25–29. https://doi.org/10.5120/ijca2018917395.

Rahmalia, Nadiyah. 2020. "Kupas Tuntas Latent Semantic Indexing Agar SEO Sukses." Glints. 25 September 2020. https://glints.com/id/lowongan/latent-semantic-indexing-adalah/.

Ramli, Fatihah, Shahrul Azman Mohd Noah, and Tri Basuki Kurniawan. 2020. "Using Ontology-Based Approach to Improved Information Retrieval Semantically for Historical Domain." International Journal on Advanced Science, Engineering and Information Technology 10 (3): 1130. https://doi.org/10.18517/ijaseit.10.3.10180.

Subramanian, Niranjan B. 2019. "Introduction to Bag of Words, N-Gram and TF-IDF | Python." AI ASPIRANT (blog). 23 September 2019. https://aiaspirant.com/bag-of-words/.

Sutabri, Tata, Agung Suryatno, Dedi Setiadi, and Edi Surya Negara. 2018. "Improving Naïve Bayes in Sentiment Analysis For Hotel Industry in Indonesia." Dalam 2018 Third International Conference on Informatics and Computing (ICIC), 1–6. Palembang, Indonesia: IEEE. https://doi.org/10.1109/IAC.2018.8780444.

Wonowidjojo, Gilbert, and Michael Sean Hartono. 2016. "Perbandingan Metode Latent Semantic Analysis, Syntactically Enhanced Latent Semantic Analysis, dan Generalized Latent Semantic Analysis dalam Klasifikasi Dokumen Berbahasa Inggris," 7.