Journal of Informatics, Electrical and Electronics Engineering
ISSN 2807-9507 (Media Online)
Vol 1, No 3, Maret 2022
Hal 97-102
https://djournals.com/jieee

Implementasi Metode Text Mining Frequency-Invers Document Frequency (Tf-Idf) Untuk Monitoring Diskusi Online

Shalvan Chamira

Program Studi Teknik Informatika Universitas Budi Darma, Medan, Indonesia
Email : Chamirashalvan@gmail.com

Abstrak-Diskusi adalah interaksi antara dua orang atau lebih. Biasanya, pertukaran ini terjadi di antara kelompok-kelompok ini dalam bentuk pengetahuan dasar atau sejenisnya, dan pada akhirnya pengetahuan ini akan memberikan pemahaman yang baik dan benar. Pembahasan bisa berupa apa saja yang semula disebut topik atau dialog. Mulai dari topik, berdiskusi dan berdiskusi, akhirnya akan membuat orang mengerti topik tersebut. Forum adalah forum atau tempat bertemunya beberapa orang yang bertujuan untuk bertukar pendapat tentang topik atau isu yang tidak terkait dengan forum tersebut. Aplikasi ini dibangun dengan menggunakan bahasa pemrograman, dan hasil pengujian pada aplikasi ini dapat digunakan sebagai metode untuk membahas keamanan data atau pesan yang dapat menjamin keamanan tanpa diketahui oleh pihak yang tidak berkepentingan

Kata Kunci: Pembelajaran; Forum Diskusi; Mobile Learning

Abstcart-Discussion is an interaction between two or more people. Usually, this exchange takes place between these groups in the form of basic knowledge or the like, and in the end this knowledge will give a good and correct understanding. Discussions can take the form of anything that was originally called a topic or dialogue. Starting from the topic, discussing and discussing, eventually it will make people understand the topic. Forum is a forum or a meeting place for several people whose aim is to exchange opinions on topics or issues that are not related to the forum. This application was built using a programming language, and the results of testing on this application can be used as a method to discuss data or message security that can guarantee security without being noticed by unauthorized parties

Keywords: Learning; Discussion Forums; Mobile Learning.

1. PENDAHULUAN

Diskusi adalah pertemuan ilmiah yang bertujuan untuk berkomunikasi dengan sekelompok orang yang membahas topik yang menjadi kepentingan umum di depan khalayak, khalayak (stasiun radio), atau khalayak (siaran TV), serta memberikan kesempatan kepada khalayak untuk bertanya dan mengutarakan pendapat (KBBI). Selain belajar di kelas, diskusi merupakan proses belajar mengajar yang wajib dilakukan pelajar khususnya mahasiswa untuk menambah wawasan, bertukar ilmu dan ide serta jajak pendapat yang mungkin jarang didapatkan di ruang kelas.

Pemanfaatan forum diskusi online khususnya dalam proses belajar mengajar merupakan salah satu trend pendidikan saat ini, dapat meningkatkan proses belajar mengajar, berbagi ilmu, pemerolehan ilmu, pembentukan ilmu dan karakter serta pembelajaran bagi peserta dan dosen. Optimalisasi proses. Guna memaksimalkan pemanfaatan dalam suatu sistem pembelajaran, maka di perlukan suatu teknik *Sistem pemantauan, sehingga informasi yang disampaikan oleh peserta dan staf pengajar dapat mencapai efek yang diinginkan*, dengan demikian istilah "Lain yang disampaikan, lain pula yang didiskusikan" dapat diatasi.

Dalam forum diskusi online terdapat beberapa masalah yang dapat terjadi, salah satu diataranya adalah ketidaksesuaian terhadap topik, hal tersebut dapat terjadi dikarenakan terlalu banyak topic dalam pembahasan. *Bahkan banyak materi yang dibahas menyimpang dari materi yang diberikan. Tentu saja hal ini memicu perdebatan di forum yang membingungkan. Debat positif* dan saling mendukung tentunya akan membentuk karakter peserta dalam menghasilkan suatu pengetahuan baru yang meningkatkan capaian hasil proses pembelajaran.

Text mining didefinisikan sebagai sebuah proses menggali informasi, dimana pengguna berinteraksi dengan dokumen-dokumen menggunakan alat analisis yang berupa komponen *data mining* yang diantaranya adalah komponen kategorisasi. *Text mining* bisa memberikan solusi atas berbagai masalah seperti *preprocessing*, pengelompokan, hingga analisa teks yang tidak terstruktur dalam jumlah yang besar. *Text mining* mengadopsi berbagai teknik dari bidang lain, seperti *Data Mining*, *Information Retrieval*, *Machine Learning*, statistik dan matematik, *linguistic*, *Natural Languange Processing* (NLP), serta *visualization*. Kegiatan terkait riset untuk *text mining* diantaranya adalah ekstraksi dan penyimpanan *text*, *preprocessing*, pengumpulan data statistik, *indexing*, dan analisis konten[1].

TF-IDF (*Term Frequency-Inverse Document Frequency*) merupakan metode statistic *numeric* yang mencerminkan seberapa pentingnya sebuah kata dalam sebuah dokumen atau korpus (Rajaraman *et al*, 2011). Hal ini sering digunakan sebagai factor bobot dalam pencarian informasi dan penambangan teks (*text mining*). Nilai *TF-IDF* meningkat secara proporsional berdasarkan jumlah atau banyaknya kata yang muncul pada dokumen, tetapi diimbangi dengan frekuensi kata dalam korpus. Variasi dari skema pembobotan*TF-IDF* sering digunakan oleh mesin pencari sebagai alat utama dalam mencetak nilai (*scoring*) dan peringkat (*ranking*) sebuah relevansi dokumen yang diberikan *user*[2].

Adapun alasan penulisan menggunakan algoritma text mining dan algrotima TF-IDF dalam menyelesaikan masalah tersebut adalah berdasarkan dari penelitian terdahulu yang dilakukan oleh Agus Putranto dengan Judul *"Perancangan Forum*

Diskusi Mobile Online Learning" menyimpulkan bahwa *materi presentasi yang diberikan berupa power point dapat dimasukkan dalam aplikasi mobile learning*. Dan penelitian yang kedua dilakukan oleh Wana Kurniawan, Andi Supranto, B. Sumardiyono dengan judul *"Rancangan Sistem forum Diskusi Online Untuk Program Studi Sistem Informasi Antara Dosen dan Mahasiswa"* Menyimpulkan bahwa merupakan salah satu forum diskusi *Memberi siswa pengetahuan umum tentang sistem informasi* secara online. Berdasarkan penjelasan diatas maka penulis menyimpulkan untuk menyelesaikan masalah tersebut pertama lakukan pengolahan teks yang dilakukan dengan algoritma text mining kemudian, selanjutkan lakukan pembobotan dari setiap kalimat tersebut dengan dengan menggunakan algoritma TF-IDF

2. METODE PENELITIAN

2.1 Text Mining

Text mining Ini dapat didefinisikan secara luas sebagai proses intensif pengetahuan di mana pengguna dapat berinteraksi dengan kumpulan dokumen dari waktu ke waktu menggunakan seperangkat alat analisis. Penambangan teks bertujuan untuk mengekstrak informasi yang berguna dari sumber data dengan mengidentifikasi dan menjelajahi pola yang menarik. Penambangan teks terutama mengarah pada pengembangan bidang penelitian data mining. Oleh karena itu, tidak mengherankan jika penambangan teks dan penambangan data akan berada pada level arsitektur yang sama [1].

2.2 Diskusi Online

Menurut "Kamus Besar Bahasa Indonesia" diskusi adalah pertemuan ilmiah. Tujuannya untuk bertukar pikiran tentang sekelompok orang yang membahas suatu masalah. Masalah tersebut merupakan topik yang menarik untuk dibahas di depan khalayak yang berkesempatan untuk bertanya atau mengutarakan pendapat. [4].

2.3 Monitoring

Monitoring adalah kegiatan yang mengamati dengan cermat suatu situasi atau situasi tertentu, termasuk perilaku atau kegiatan tertentu, untuk memperoleh semua masukan data dari pengamatan tersebut. Atau informasi. Menjadi dasar untuk menentukan langkah-langkah selanjutnya yang diperlukan.

2.4 Term Frequency- Inverce Document Frequency (TF-IDF)

Metode TF-IDF menggabungkan dua konsep, yaitu frekuensi kata dalam dokumen dan pengacakan dokumen yang berisi kata [6]. Saat menggunakan TF-IDF untuk menghitung bobot, pertama-tama hitung nilai TF dari kata tersebut, dan bobot setiap kata adalah 1. Dan nilai IDF diwakili oleh persamaan (1).

$$IDF(word) = log\,\frac{td}{df} \quad (1)$$

IDF (kata) adalah nilai IDF dari setiap kata yang akan dicari, td adalah jumlah dokumen yang ada, dan df adalah berapa kali kata tersebut muncul di semua dokumen.

Algoritma MMR adalah metode untuk meringkas satu dokumen dan banyak dokumen, dan efektif dan sederhana. Ringkas dokumen ekstraktif, nilai akhir dihitung dengan rumus dalam kalimat di MMR:

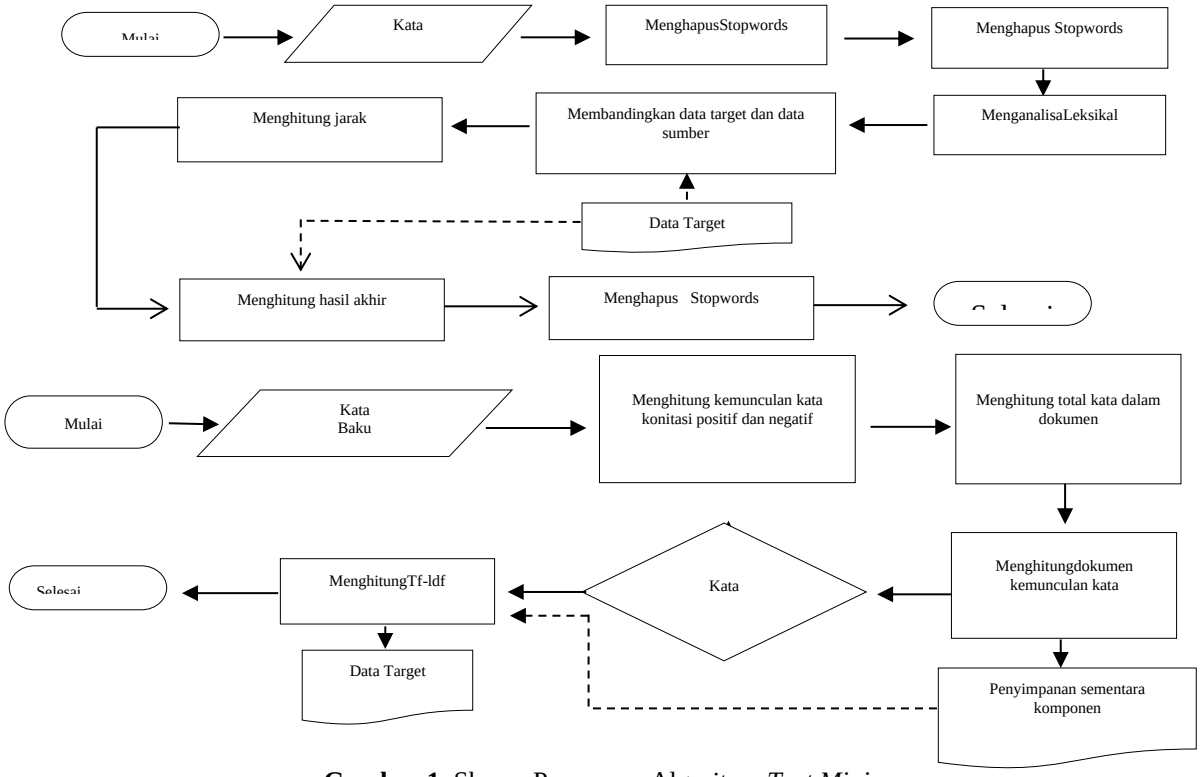
MMR = argmax[ג* $Sim_1(D, Q) -$ $(1 -$ ג) * max $Sim_2(D_i, D')$]

Nilai parameter λ berkisar dari 0 hingga 1 atau 0,1. Jika parameter λ = 1, nilai MMR yang diperoleh akan cenderung berhubungan dengan dokumen aslinya. Ketika nilai parameter λ = 0, maka nilai yang diperoleh cenderung berhubungan dengan kalimat yang telah digali sebelumnya, yang akan dibandingkan. Untuk meringkas, gunakan nilai parameter λ = 0.7 atau λ = 0.8 untuk dokumen pendek.

3. HASIL DAN PEMBAHASAN

Algoritma term frequency inverse document frequency (TF-IDF) merupakan metode untuk menentukan jarak antara sebuah kata (term) dan sebuah dokumen dengan memberikan bobot pada tiap kata. Penerapan dari kedua algoritma tersebut dibagi menjadi algoritma text mining digunakan sebagai penolahan teks dari setiap kalimat, hasil dari algoritma tersebut akan menghasilkan kata-kata akar (root) dari setiap kalimat. Kata-kata akar tersebut akan dibandingkan dengan kata-kata yang mengandung konotasi positif dan konotasi negatif dengan menggunakan algoritma *Term Frequency Invers Document Frequency* (TF-IDF). Hasil dari algoritma TF-IDF berupa bobot kesesuaian atas kata-kata konotasi negatif dan positif yang telah ditentukan. Penerapan dari algoritma *text mining* dan algoritma *Term Frequency Invers Document Frequency* (TF-IDF) dapat dilihat pada gambar 1.

ISSN 2807-9507 (Media Online)
Vol 1, No 3, Maret 2022
Hal 97-102
https://djournals.com/jieee



Gambar 1. Skema Penerapan Algoritma *Text Mining*

Tahap analisis dilakukan dengan menganalisis metode mesin temu kembali informasi, yang meliputi analisis preprocessing teks pada kumpulan dokumen (korpus), yaitu: resolusi kalimat, pelipatan kasus, pemfilteran kalimat, analisis penandaan kata dan stemming. Selain itu, algoritma TF-IDF dianalisis melalui serangkaian kegiatan untuk melakukan operasi berikut: menghitung jumlah kata dalam kalimat, menghitung jumlah kata dalam dokumen, menghitung nilai frekuensi terbalik dokumen, menghitung nilai bobot kata, dan Hitung nilai kumulatif W untuk setiap kalimat.

Untuk menganalisa kalimat atas berdiskusinya, maka langkah awal yang dilakukan penulis adalah mengumpulkan data kalimat dari suatu situs *forum-forum online*, dalam hal ini penulis memanfaatkan teknik web data extraction. Untuk memudahkan dalam melakukan penerapan algoritma tersebut dalam melakukan analisa kalimat. Kalimat pada dokumen dipecahkan menjadi 9 kalimat sesuai dengan aturan pemecahan. Hasilnya dapat dilihat pada Gambar 2.

> Stopword adalah kata umum (common words) yang biasanya muncul dalam jumlah besar dan tidak memiliki makna. Stopword umumnya dimanfaatkan dalam task informasi retrieval, termasuk oleh google. Contoh stopword untuk bahasa inggris diantaranya of dan the. Sedangkan untuk bahasa Indonesia diantaranya yang, di dan ke.
> Saya pernah membuat stopword bahasa Indonesia untuk tugas salah satu mata kuliah. Tujuannya waktu bukan untuk information retrieval, tapi untuk klasifikasi. Saya gunakan stopword untuk mengurangi jumlah kata yang harus diproses.
> Saya membuat daftar stopword dengan cara mengumpulkan kata paling banyak muncul pada korpus (saya menggunakan berita kompas), setelah diurutkan kemudian diperiksa secara manual satu persatu. Karena daftar itu dibuat secara manual dan untuk task klasifikasi, ada beberapa kata yang mungkin di perdebatkan, jadi silahkan edit sesuai kebutuhan.

Gambar 2. Data Kalimat

Setelah data dikumpulkan maka tahapan selanjutnya dari algoritma *text mining* adalah sebagai berikut :

a. Case Folding

Adapun tahapan yang dilakukan pada algoritma *Teks Mining* diawali dengan melakukan proses *Case Folding*, yaitu melakukan perubahan string menjadi huruf kecil. Hasil dari proses *case folding* dapat dilihat pada paragraf dibawah ini.

ISSN 2807-9507 (Media Online)
Vol 1, No 3, Maret 2022
Hal 97-102
https://djournals.com/jieee

Tabel 1. Data Kalimat

No.	Kalimat
Q	Stopword untuk bahasa Indonesia
1.	stopword adalah kata umum common words yang biasanya muncul dalam jumlah besar dan tidak memiliki makna
2.	stopword umumnya dimanfaatkan dalam task informasi retrieval termasuk oleh google
3.	contoh stopword untuk bahasa inggris diantaranya of dan the
4.	sedangkan untuk bahasa indonesia diantaranya yang di dan ke
5.	saya pernah membuat stopword bahasa indonesia untuk tugas salah satu mata kuliah
6.	tujuannya waktu bukan untuk information retrieval tapi untuk klasifikasi
7.	saya gunakan stopword untuk mengurangi jumlah kata yang harus diproses
8.	saya membuat daftar stopword dengan cara mengumpulkan kata paling banyak muncul pada korpus saya menggunakan berita kompas setelah diurutkan kemudian diperiksa secara manual satu persatu
9.	karena daftar itu dibuat secara manual dan untuk task klasifikasi ada beberapa kata yang mungkin di perdebatkan jadi silahkan edit sesuai kebutuhan

b. Filtering

Pada tahapan ini akan dilakukan pengahapusan kata yang tidak penting atau disebut dengan stop word, dalam hal ini penulis memanfaatkan library tala, beberapa daftar kata stop word sebagai berikut : *ada, adalah, adanya, adapun, akan, akankah, akhir, akhiri, akhirnya, aku, akulah, amat, amatlah, anda, andalah, antar, antara, dll.* Hasil dari menghilangkan kata stop word dari berita dapat dilihat pada paragraf dibawah ini :

Tabel 2. Hasil Tahapan *Case Folding*

No.	Kalimat
Q	Stopword untuk bahasa Indonesia
1.	stopword common words makna
2.	stopword umumnya dimanfaatkan taskinformasi retrieval google
3.	contoh stopword bahasa inggris diantaranya
4.	bahasa indonesia diantaranya
5.	stopword bahasa indonesia tugas mata kuliah
6.	tujuannya information retrieval klasifikasi
7.	gunakan stopword mengurangi diproses
8.	daftar stopword mengumpulkan korpus berita kompas diurutkan diperiksa manual per
9.	daftar manual task klasifikasi di perdebatkan silahkan edit

c. *Tokenizing*

Tahap *tokenizing* adalah tahap pemotongan string menjadi potongan kata kemudian disusun menjadi baris. Tahap tersebut dapat dilihat dibawah ini.

Tabel 3. Hasil *filtering* kata

Kata	Kata	Kata	Kata
stopword	google	mengumpulkan	Edit
diperdebatkan	contoh	berita	umumnya
bahasa	inggris	kompas	diantaranya
indonesia	tugas	diurutkan	per
common	mata kuliah	diperiksa	diproses
makna	tujuannya	manual	retrieval
dimanfaatkan	klasifikasi	korpus	daftar
task	gunakan	dibuat	silahkan
information	mengurangi	words	-

d. *Stemming*

Pada tahapan dilakukan proses mentransformasikan kata-kata yang menggunakan aturan tertentu seperti menghilangkan awalan dan akhiran untuk mendapatkan kata-kata akarnya atau disebut dengan root. Hasil dari proses stemming dapat dilihat pada paragraf dibawah ini :

Tabel 4. *Tokenizing* Kata

Kata	Kata	Kata	Kata
antara	google	kurang	silah
bahasa	guna	makna	stopword
berita	information	manfaat	task
buat	indonesia	manual	tugas
common	inggris	mata kuliah	tujuan
contoh	klasifikasi	per	umum
daftar	kompas	periksa	urut
debat	korpus	proses	words
inf	kumpul	retrieval	-

Pembobotan dilakukan dengan menggunakan proses TF dan IDF dan perangkingan. Dengan menggunakan metode *cosine similarity* nilai bobot *relevance query* dan bobot *similarity* diperoleh. Hasil yang diperoleh dapat dilihat pada Tabel berikut ini:

Tabel 5. Hasil Tahapan *Stemming*

	D1	D2	D3	D4	D5	D6	D7	D8	D9
D1	1	0.009	0.012	0	0.012	0	0.011	0.011	0
D2	0.009	1	0.010	0	0.010	0.287	0.009	0.006	0.092
D3	0.012	0.010	1	0.398	0.103	0	0.012	0.007	0
D4	0	0	0.398	1	0	0	0	0	0
D5	0.012	0.010	0.103	0.398	1	0	0.012	0.007	0
D6	0	0.287	0	0	0	1	0	0	0.125
D7	0.011	0.009	0.012	0	0.012	0	1	0.007	0
D8	0.007	0.006	0.007	0	0.007	0	0.007	1	0.137
D9	0	0.092	0	0	0	0.125	0	0.137	1

Pada penelitian ini MMR menggunakan nilai parameter λ = 0.7[3]. Hasil iterasi dan bobot MMR dapat dilihat pada Gambar berikut :

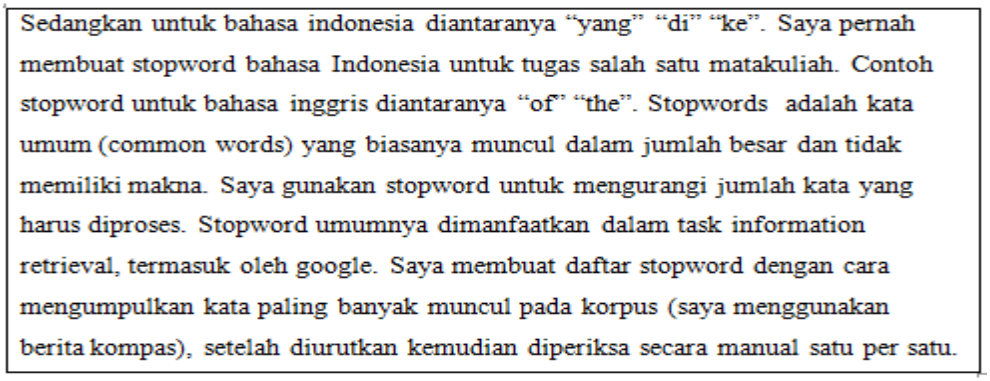
Tabel 6. Hasil *Similarity* antar kalimat pada *Postingan*

iterasi ke	D1	D2	D3	D4	D5	D6	D7	D8	D9
1	0.016	0.013	0.138	**0.532**	0.366	0	0.016	0.009	0
2	0.016	0.013	0.019	-	**0.247**	0	0.016	0.009	0
3	0.016	0.013	**0.019**	-	-	0	0.016	0.009	0
4	**0.012**	0.010	-	-	-	0	0.012	0.008	0
5	-	0.010	-	-	-	0	**0.012**	0.008	0
6	-	**0.010**	-	-	-	0	-	0.008	0
7	-	-	-	-	-	-0.086	-	**0.008**	-0.027

Tabel 7. Hasil iterasi MMR

Iterasi ke	Kalimat	Bobot ArgMax MMR
MMRMAX1	D4	**0.532**
MMRMAX2	D5	**0.247**
MMRMAX3	D3	**0.019**
MMRMAX4	D1	**0.012**
MMRMAX5	D7	**0.012**
MMRMAX6	D2	**0.010**
MMRMAX7	D8	**0.008**

Output hasil ringkasan akhir dapat dilihat pada Gambar 3.8. Proses TF-IDF yang dilakukan pada dokumen postingan juga diberlakukan pada dokumen komentar peserta. Selanjutnya perbandingan antara dokumen tersebut dilakukan untuk menentukan kelayakan komentar.


Sedangkan untuk bahasa indonesia diantaranya "yang" "di" "ke". Saya pernah membuat stopword bahasa Indonesia untuk tugas salah satu matakuliah. Contoh stopword untuk bahasa inggris diantaranya "of" "the". Stopwords adalah kata umum (common words) yang biasanya muncul dalam jumlah besar dan tidak memiliki makna. Saya gunakan stopword untuk mengurangi jumlah kata yang harus diproses. Stopword umumnya dimanfaatkan dalam task information retrieval, termasuk oleh google. Saya membuat daftar stopword dengan cara mengumpulkan kata paling banyak muncul pada korpus (saya menggunakan berita kompas), setelah diurutkan kemudian diperiksa secara manual satu per satu.

Gambar 8. Hasil ringkasan

4. KESIMPULAN

Dari hasil penelitian diperoleh beberapa kesimpulan diantaranya adalah:

a. Metode TF-IDF, cosine similarity dan MMR telah berhasil diterapkan pada peringkasan dokumen untuk memantau diskusi online ini.
b. Pada menerapkan nilai MMR untuk menentukan kualifikasi komentar ini, dapat dipertahankan dalam forum diskusi online bahwa telah berhasil diterapkan dan dapat digunakan oleh semua orang.
c. Hasil pengujian yang diberikan pada black box dan hasil UAT dapat disimpulkan bahwa sistem sudah layak dan dapat digunakan untuk membantu dalam memantau proses diskusi online.

REFERENCES

[1]. Kurniwan,B; Efendi,S;& Sitompul, O, S. (2012). Klasifikasi Konten Berita Dengan Metode Text Mining. Jurnal Teknologi Informasi, 14-19
[2]. Rizki, Dhidik, dan Eko Supraptono. 2017. Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen. Skripsi. Jurusan Teknik Elektro. Fakultas Teknik. Universitas Negeri Semarang Kampus Sekaran, Gunungpati, Semarang
[3]. Tata Sutabri, *Analisis Sistem Informasi*. Yogyakarta: Andi, 2012
[4]. Kamus Besar Bahasa Indonesia (KBBI)
[5]. Kusrin, Emha Taufiq Luthfi,(2009). *Algoritma Data Mining*. Yogyakarta: ANDI.
[6]. Abdul Kadir, "Algoritma & Pemrograman Menggunakan Java", Yogyakarta: Andi, 2012
[7]. Soeherman, Bonnie dan Marion Pinontoan. (2008). Designing Information System. Jakarta : PT. Elex Media Komputindo.
[8]. A. Firman, H. F. Wowor, X. Najoan, J. Teknik, E. Fakultas, and T. Unsrat, "Sistem Informasi Perpustakaan Online Berbasis Web," vol. 5, no. 2, 2016.
[9]. R. S. dan J. Febio, "Membangun Aplikasi E-Library Menggunakan Html, Php Script, Dan Mysql Database,
[10]. N. R. YANTI and 14110615, "Implementasi Algoritma Camellia Pada Penyandian Record Database," 2019.
[11]. N. B. Batubara and 5110695, "Implementasi Metode Even-Rodeh Code Untuk Kompresi Kitab Undang-Undang Hukum Pidana (Kuhp) Berbasis Android," 2019.
[12]. N. B. Batubara and 5110695, "Implementasi Metode Even-Rodeh Code Untuk Kompresi Kitab Undang-Undang Hukum Pidana (Kuhp) Berbasis Android," 2019.
[13]. P. Studi, I. Komputer, and F. U. Mulawarman, "Memahami Penggunaan UML (Unified Modelling Language)," vol. 6, no. 1, pp. 1–15, 2011.
[14]. A. Firman, H. F. Wowor, X. Najoan, J. Teknik, E. Fakultas, and T. Unsrat, "Sistem Informasi Perpustakaan Online Berbasis Web," vol. 5, no. 2, 22