# ANALISIS TINGKAT PLAGIASI DOKUMEN SKRIPSI DENGAN METODE COSINE SIMILARITY DAN PEMBOBOTAN TF-IDF

*(ANALYSIS OF THESIS PLAGIARISM DOCUMENT LEVEL WITH COSINE SIMILARITY METHOD AND TF-IDF WEIGHTING)*

**Muhammad Azmi**[1)]
[1)]Prodi Sistem Informasi STMIK Syaikh Zainuddin NW Anjani
Jalan Raya Mataram-Lb. Lombok, KM 49 Anjani, Lombok Timur-NTB, Indonesia
e-mail: muhammad.azmi@stmiksznw.ac.id[1)]

**ABSTRAK**

*Plagiarisme merupakan kegiatan menduplikasi atau meniru hasil karya orang lain kemudian diakui sebagai karyanya sendiri tanpa seizin penulisnya atau mencantumkan sumbernya. Plagiarisme atau Penjiplakan bukanlah sesuatu yang sulit untuk dilakukan karena dengan menggunakan teknik copy-paste-modify sebagian ataupun keseluruhan dokumen tersebut maka dokumen tersebut sudah bisa dikatakan hasil plagiat atau duplikasi.*

*Praktek plagiarisme terjadi akibat mahasiswa sudah terbiasa mengambil tulisan orang lain tanpa mencantumkan sumber asalnya, bahkan menyalin secara keseluruhan dan sama persis.*

*Salah satu cara yang bisa digunakan untuk mencegah praktek plagiarisme yaitu dengan melakukan pencegahan dan mendeteksi. Pendeteksian plagiarisme menggunakan konsep similarity atau kemiripan dokumen merupakan salah satu cara untuk mendeteksi copy & paste plagiarism dan disguised plagiarism. salah satu metode yang tepat yang bisa dilakukan untuk mendeteksi plagiarisme dengan menganalisis tingkat plagiarisme dokumen menggunakan metode Cosine Similarity dan Pembobotan TF-IDF.*

*Penelitian ini menghasilkan sebuah aplikasi yang mampu memproses nilai kemiripan dokumen yang akan diuji. Hasik pengujian menunjukan sudah sesuai antar perhitungan manual dan implementasi algoritma dalam aplikasi yang dibuat. Penggunaan Library Sastrawi Cukup Efektif dalam proses Stemming. Perhitungan yang menggunakan stemming akan memiliki nilai kemiripan yang lebih tinggi dibandingan dengan perhitungan tanpa metode stemming.*

***Kata Kunci**: Plagiarisme, Consine Similarity, Pembobotan TD-IDF.*

**ABSTRACT**

*Plagiarism is the activity of duplicating or imitating the work of others and then being recognized as their own work without the permission of the author or citing the source. Plagiarism or plagiarism is not something that is difficult to do because by using the copy-paste-modify technique of part or all of the document, the document can be said to be the result of plagiarism or duplication. The practice of plagiarism occurs because students are accustomed to taking other people's writings without including the original source, even copying them in their entirety and in the exact same way.*

*One way that can be used to prevent plagiarism is to prevent and detect. Plagiarism detection using the concept of similarity or document similarity is one way to detect copy & paste plagiarism and disguised plagiarism. one of the appropriate methods that can be done to detect plagiarism is to analyze the level of document plagiarism using the Cosine Similarity method and TF-IDF Weighting.*

*This research produces an application that is able to process the similarity value of the document to be tested. The test results show that it is appropriate between manual calculations and the implementation of the algorithm in the application made. The Use of Literary Libraries is Quite Effective in the Stemming Process. Calculations using stemming will have a higher similarity value than calculations without the stemming method.*

***Keywords**: Plagiarism, Cosine Similarity, Weighting TF-IDF*

## I. PENDAHULUAN

Perkembangan teknologi informasi terutama teknologi digital diera modern saat ini sudah banyak dimanfaatkan oleh masyarakat dan sudah menjadi kebutuhan yang tidak bisa dipisahkan dari kehidupan manuisa modern. Salah satu komponen penting dalam didalam dunia digital adalah dokumen. Dokumen yang sudah berbentuk digital memudahkan dalam hal penyimpanan, mudah dalam pencarian ,efisien dan bahkan mudah untuk di duplikat atau dijiplak. Penjiplakan atau palgiarisme sudah sangat sering terjadi terutama dalam dunia akademik, dari tingkat sekolah bahkan sampai tingkat perguruan tinggi. Praktek penjiplakan banyak dilakukan untuk menyelesaikan tugas dengan mudah dengan teknik *copy-paste-modify* tanpa perlu mempelajari atau mengekplorisasi materinya, bahkan penjilakan banyak dilakukan mahasiswa terutama saat sedang menyelesaikan tugas akhir atau skripsi.

Plagiarisme merupakan kegiatan menduplikasi atau meniru hasil karya orang lain kemudian diakui sebagai karyanya sendiri tanpa seizin penulisnya atau mencantumkan sumbernya. Plagiarisme atau Penjiplakan bukanlah sesuatu yang sulit untuk dilakukan karena dengan menggunakan teknik *copy-paste-modify* sebagian ataupun keseluruhan dokumen tersebut maka dokumen tersebut sudah bisa dikatakan hasil plagiat atau duplikasi [1]. plagiarisme sendiri dapat dikelompokkan menjadi beberapa kelompok berdasarkan proporsi atau presentase kata, kalimat atau paragraf yang diduplikat yaitu, pagiarisme ringan (<30%), plagiarisme sedang (30- 70%) dan plagiarisme berat (>70%)[2].

Praktek plagiarisme terjadi akibat mahasiswa sudah terbiasa mengambil tulisan orang lain tanpa mencantumkan sumber asalnya, bahkan menyalin secara keseluruhan dan sama persis. Praktek plagiarisme banyak dilakukan mahasiswa terutama saat menyelesaikan tugas akhir atau skripsi.

Salah satu cara untuk mencegah praktek plagiarisme atau plagiat yaitu dengan melakukan pencegahan dan melakukan pendeteksian dini. Mencegah berarti menjaga dan mencegah supaya praktek palgiarisme tidak dilakukan. Usaha ini harus dilakukan terutama pada sistem pendidikan. Salah satu cara yang bisa dilakukan untuk untuk mendeteksi dokumen plagiat dengan cara melakukan perbandingan dokumen secara manual denga melakukan pencocokan terhadap sebuah dokumen, hal tesebut dinilai kurang efektif dan efisien.

Pendeteksian plagiarisme atau penjiplakan menggunakan konsep *similarity* atau kemiripan dokumen merupakan cara untuk mendeteksi *copy & paste plagiarism dan disguised plagiarism* [3]. salah satu metode yang tepat yang bisa dilakukan untuk mendeteksi plagiarisme dengan menganalisis tingkat plagiarisme dokumen menggunakan metode *Cosine Similarity* dan Pe bobotan TF-IDF.

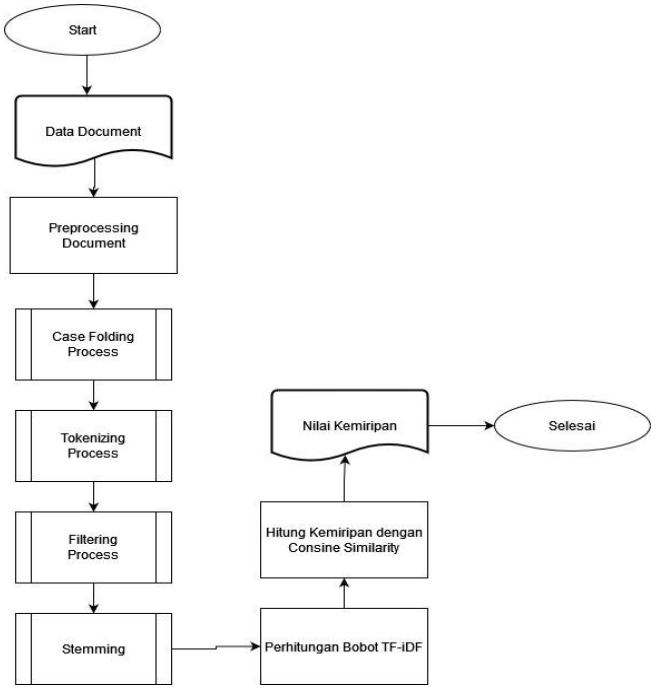
Berdasarkan latar belakang diatas dan mengacu kepada penelitian sebelumnya, maka dibuatlah penelitian ini dengan menggunakan metode *Cosine Similarity* dan Pembobotan TF-IDF untuk mendeteksi kemiripan dokumen skripsi. Selanjutnya berdasarkan hasil analisisnya maka akan dibahas bagaimana motode *Consine Similarity* dan Pembobotan TF-IDF di implementasikan pada sebuah aplikasi dalam mendeteksi plagiat dokumen skripsi.

## II. STUDI PUSTAKA

Pembuatan Aplikasi atau penelitian ini mengacu pada penelitian- penelitian sebelumnya seperti yang dilakukan Salmuasih dan Andi Sunyoto membahas tentang bagaimana implementasi *Algoritma Rabin Karp* untuk pendeteksian Plagiat dokumen menggunakan konsep *similarity*, Fitri Dwi indah kusuma dkk (2015) melakukan penelitian dengan melakukan Aplikasi Pendeteksi Kemiripan Laporan menggunakan *teks mining* dan *Clustering*, kemudian Lasmedi Afuan juga melakukan penelitian tentang *Stemming dokumen* teks bahasa indonesia menggunakan Algoritma *Porter, Tinalia*h dan Triana Elizabeth dengan penelitian yang berjudul Perbandingan Hasil Deteksi Plagiarisme Dokumen dengan metode *Jaro-Winkler Diistance* dan *Metode Latent Semantic Analys.*

## III. METODE PENELITIAN

Desain Alur Penelitian merupakan gambaran umum untuk menentukan alur atau langkah demi langkah yang akan dilakukan dalam penelitian ini dilihat pada Gambar 1.

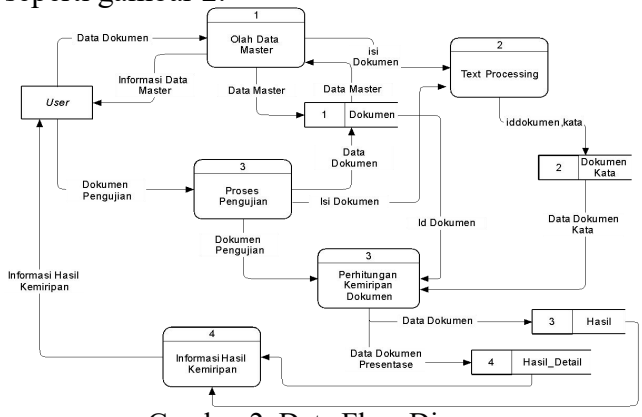



Gambar 1. Desain Penelitian

## IV. HASIL DAN PEMBAHASAN

### *A.* Desain Sistem dengan *Flow Diagram*

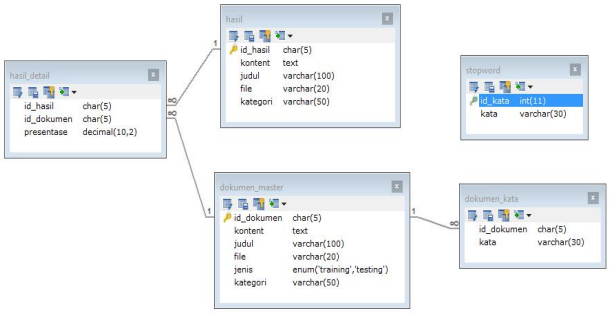
Perancangan proses pada penelitian ini yatu *Data Flow Diagram*. DFD menggambarkan aliran data yang didalam sistem, apa yang menjadi inputan, proses yang terjadi dalam system serta output yang dihasilkan oleh sistem. Pada DFD menggambarkan aktivitas yang terdiri dari sisi pengguna (*user*) maupun sistem yang ditunjukan seperti gambar 2.




Gambar 2. Data Flow Diagram

### *B.* Perancangan Basis Data

Basis data diperlukan untuk memberikan gambaran tentang media penyimpanan data yang menghasilkan informasi. Adapun desain basis data berupa Relasi Antar Tabel yang ditunjukan dalam gambar 3 dengan 5 tabel yang meliputi tabel dokemen_master, dokumen_kata, hasil, hasil_detail dan stopword.

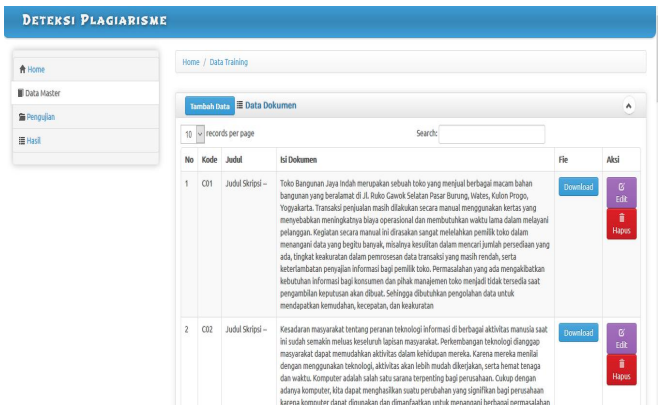



Gambar 3. Struktur Basis Data

### *C.* Pengujian Sistem

Tahapan pengujian dilakukan untuk memastikan semua menu yang ada pada sistem ini dapat berjalan dan berfungsi dengan baik.pengujian dilakukan dengan memasukkan dokumen latar belakang skripsi dan dokumen latar belakang yang sudah di setujui sebagai pembanding, dilakukan proses *stemming* dan *no stemming,* melakukan uji deteksi kemiripan dokumen, melakukan perbandingan banyak dokumen, dan melakukan perbandingan dengan sistem serupa untuk menguji akurasi dan presisi hasil deteksi dokumen.

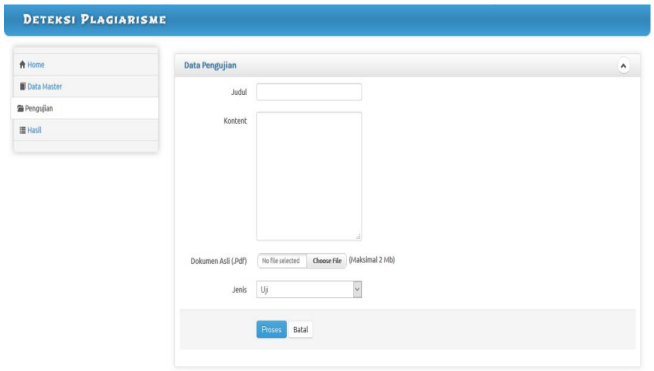
1) Tampilan Home atau data master merupakan tampilan antarmuka yang muncul pertama kali saat aplikasi dijalankan.




Gambar 4. Dashboard Awal

2) Menu yang akan yang diuji adalah menu pengujian, dimana dihalaman ini dilakukan perbandingan. Tahapan yang dapat dilakukan pada menu ini adalah memasukan dua dokumen yaitu dokumen asli dan dokumen

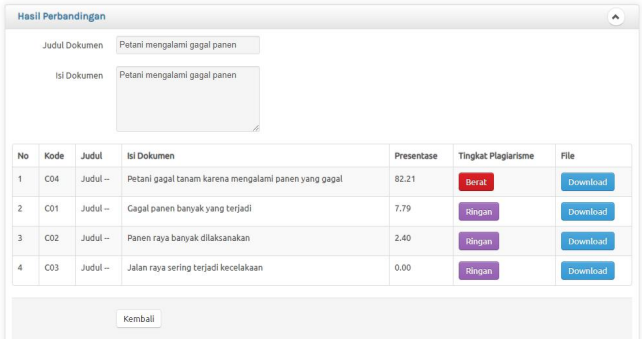
pembanding yang diupload dan akan diperiksa ke dalam text editor.




Gambar 5. Implemenayasi Halaman Pengujian

## *D.* Hasil Analisa

Pada pengujian sistem yang dilakukan, sistem memeriksa dua dokumen yaitu dokumen asli dan dokumen pembanding. Dilakukan pengujian dengan menguji perhitungan metode cosine similarity dan pembobotan *TF-IDF*. Berikut hasil pengujian dengan aplikasi yang dibuat yangditunjukan dalam gambar 6.




Gambar 6. Pengujian Perhitungan Hasil

Setelah dilakukan pengujian pada aplikasi dengan perhitungan metode cosine similarity dan pembobota TF-IDF, maka sistem dapat dibandingkan dengan perhitungan manual dengan data yang ada sehingga didapatkan hasil seperti dibawaah ini.

$$\text{Cosine D1} = \frac{Q\ X\ D1}{|Q|\ X\ |D1|} = \frac{0{,}03}{|0{,}86|\ x\ |0{,}45|} = \frac{0{,}03}{0{,}387} = 0{,}077$$

$$\text{Cosine D2} = \frac{Q\ X\ D2}{|Q|\ X\ |D2|} = \frac{0{,}015}{|0{,}86|\ x\ |0{,}74|} = \frac{0{,}015}{0{,}63} = 0{,}02$$

$$\text{Cosine D3} = \frac{Q\ X\ D3}{|Q|\ X\ |D3|} = \frac{0}{|0{,}86|\ x\ |0{,}95|} = 0$$

$$\text{Cosine D4} = \frac{Q\ X\ D4}{|Q|\ X\ |D4|} = \frac{0{,}769}{|0{,}86|\ x\ |1{,}07|} = \frac{0{,}769}{0{,}920} = 0{,}83$$

Dari hasil pengujian yang dilakukan didapatkan nilai yang sama antara pengujian yangdilakukan oleh sistem dengan perhitungan manual. Pengujian selanjutnya dilakukan dengan perhitungan manual yang sama dengan data yang berbeda seperti asil perhitungan manual dibawah ini :

$$\text{Cosine D1} = \frac{Q\ X\ D1}{|Q|\ X\ |D1|} = \frac{5{,}783}{|2{,}880|\ x\ |4{,}075|} = \frac{5{,}783}{11{,}737} = 0{,}493$$

$$\text{Cosine D2} = \frac{Q\ X\ D2}{|Q|\ X\ |D2|} = \frac{2{,}980}{|0{,}86|\ x\ |0{,}74|} = \frac{2{,}980}{12{,}519} = 0{,}238$$

Dari hasil perhitungan manual didapatkan hasil di aplikasi pada gambar 7.


Gambar 7. Pengujian Perhitungan Kedua

Dari kedua pengujian hitung manual maka dapat dikatakan aplikasi yang dikembangkan sudah sesuai antara perhitungan manual dengan penerpan algoritma dalam aplikasi yang dibangun.

Setelah dilakukan pengujian sederhana seperti diatas, selanjutnya dilakukan pengujian dengan Stemming dan tanpa Stemmin yang ditunjulan dalam tabel 1:

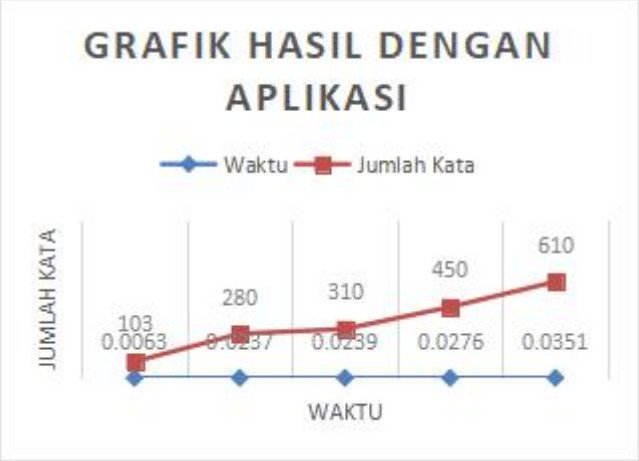
Tabel 1. Perbandingan Tingkat *Plagiarisme* dengan *Stemming* dan Tanpa *Stemming*

| No | Dokumen | Dengan Stemming | Tanpa Stemming |
|---|---|---|---|
| 1 | D5 | Sedang | Ringan |
| 2 | D6 | Sedang | Ringan |
| 3 | D7 | Ringan | Ringan |
| 4 | D8 | Sedang | Sedang |
| 5 | D9 | Ringan | Sedang |
| 6 | D10 | Sedang | Sedang |
| 7 | D11 | Sedang | Sedang |
| | Jumlah Ringan | 2 | 3 |
| | Jumlah Sedang | 5 | 4 |

Dari tabel 1 dapat dilihat hasil pengujian kemiripan pada tahapan yang menggunakan

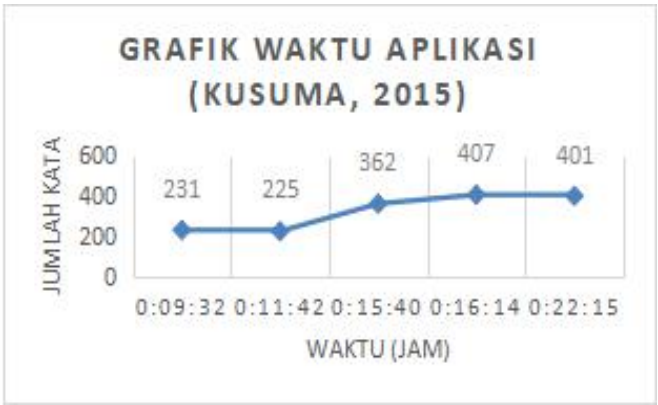
*stemming* didapatkan hasil nilai kmeiripan sedang 5 dan kemiripan ringan 2 begitu juga dengan tanpa steming yang didpatkan hasil jumlah ringan 3 dan jumlah sedang 4. Hasill dengan stemming menghasilkan nilai kemiripan yang lebih besar dibadningan tanoa tahaopan stemming.

Pada tahapan ini dilakukan pengujian dengan membandingkan aplikasi yang dibangun dengan aplikasi yang serupa, aspek yang dibandingkan antara lain tingkat presisi dan waktu memrosesan dokumen. Pengujian ini untuk menguji *library* yang digunakan dalam penelitian in yaitu dengan *library sastrawi* dengan meniadakan aspek komponen *hardware* atau teknologi yang digunakan. Hasil pengujian tersebut divisualkan dalam gambar 8:




Gambar 8. Grafik Waktu Aplikasi

Dari gambar 8 dapat dilihat pengujian yang dilakukan menunjukan semakin banyak kata yang diproses maka waktu yang dibutuhkan semakian lama. Pada jumla kata dengan 103 didapatkan nilai waktu 0,0063 detik sebanding dengan jumlah kaya 610 dengan waktu 0,0351 detik. Pengujian yang sama dilakukan oleh (Kusuma, 2015) yang dapat dilihat dalam gambar 9.




Gambar 9. Grafik Aplikasi (Kusuma, 2015)

Hasil perbandingan dari gamabr 8 dan 9 menujukan rentang waktu yang cukup jauh sebagai contoh data denga jumlah kata 231 kata membutuhkan waktu 9 detik lebih, sedangkan dalam aplikasi yang dibangun oleh penulis kata dengan 610 tidak sampai 1 detik waktu yang dibutuhkan.

## V. KESIMPULAN

Berdasarkan hasil analisia dan pengujian yang telah penulis lakukan, maka dapat ditarik kesimpulan sebagai berikut:

1. Tahapan stemming membutuhkan waktu yang semakin lama ketika kata yang diproses semakin banyak.
2. Hasil pengujian sistem yang dilakukan dengan perhitungan manual dengan implementasi algoritma sudah sesuai hasil yang diharapakan. Hal ini ditunjukan dari 2 pengujian yang dilakukan dengan dokumen yang berbeda menghasilkan output keluaran yang sama.
3. Penggunaan *Stemming* dalam tahapan *preprocessing* dokumen mampu meningkatkan sensistifitas kemiripan dokumen dengan dibuktikan kecenderungan nilai kemiripan yang lebih tinggi dibanding pemroses tanpa menggunakan stemming.

## DAFTAR PUSTAKA

[1] Irianto, WA., 2014, *Penentuan Tingkat Plagiarisme Dokumen Penelitian Menggunakan Cenroid Lingkage Hierarchical Method (Clhm),* Jurnal Program Teknologi Informasi Dan Ilmu Komputer. Universitas Brawijaya Malang.

[2] S. Sastroasmoro., 2006, *Beberapa catatan tentang, Majalah Kedokteran Indonesia*, Vol. 55, Hal. 1.

[3] Salmuasih., Sunyoto Andi., 2013, *Implementasi Algoritma Rabin Karp untuk pendeteksian Plagiat Dokumen Teks Menggunakan Konsep Similarity*, Seminar Nasional Aplikasi Teknologi Informasi (SNATI), Yogyakarta.

[4] Pemerintah Indonesia. 2010. *Peraturan Mendiknas Republik Indonesia No. 17 Tahun 2010 Tentang Pencegahan dan Penanggulangan Plagiat di Perguruan Tinggi*. Lembaran Negara RI Tahun 2010. Kemendikbud. Jakarta.

[5] Qaiser, Shahzad., 2018, *Text Mining: Use of TF-IDF to Examine the Relevance of*

*Words to Documents*, International Journal of Computer Applications (0975 – 8887) Volume 181 – No.1, July 2018.

[6] H. Wu and R. Luk and K. Wong and K. Kwok., 2008, *Interpreting TF-IDF term weights as making relevance decisions*, ACM Transactions on Information Systems, 26 (3).