PROPOSAL TUGAS AKHIR

Pengujian Deteksi Tingkat Plagiarisme dengan Mempertimbangkan Struktur Kalimat dan Makna dengan Menggunakan Algoritma *Decision Tree*

Oleh:

DENNY GUNAWAN
NIM. 121111362

RIVALDI WARMAN
NIM. 121113223

WISELY JANSEN HADI
NIM. 121111044

Disetujui Oleh:

Dosen Pembimbing I,	Dosen Pembimbing II,
Dr. Pahala Sirait, S.T., M.Kom.	**Sunario Megawan, S.Kom.,M.kom,**
NIP. 45970117	**NIP. 45061039**

PROGRAM STUDI TEKNIK INFORMATIKA

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER

MIKROSKIL

MEDAN

2016

1 Latar Belakang

Permasalahan terhadap plagiarisme telah ada selama berabad-abad. Namun penyebarluasan informasi dan teknologi komunikasi, termasuk internet telah memberikan kontribusi untuk kemudahan menjiplak. Banyak layanan *online* yang ada untuk memfasilitasi plagiarisme, termasuk *database* esai dan alat teks "*synonymizer*" seperti synimizer.com yang hasil keluaran dari masukan teks berupa daftar-daftar sinomin dari tiap kata. Hasil survei dari 80.000 lebih mahasiswa di amerika serikat dan kanada menunjukkan bahwa sebanyak 62% sarjana dan 59% pascasarjana melakukan plagiarisme *cut and paste* dari sumber tertulis dan internet (McCabe, 2005). Di Indonesia sendiri banyak terjadi kasus plagiarisme salah satunya yaitu dosen di fakultas ekonomi dan bisnis (FEB) Universitas Gadjah Mada, Anggito Abimanyu terhadap artikelnya "Gagasan Asuransi Becana" yang terbit di harian kompas. 10 Februari 2014. Tulisan ini memiliki kesamaan dengan artikel Hotbonar Sinaga dan Munawar Kasan yang Berjudul "Menggagas Asuransi Becana". (Lestarini, Ade Hapsari, 25 Februari 2014)

Penelitian pada sistem untuk menganalisa dan mendeteksi plagiarisme telah dilakukan hampir dua dekade. Sistem pendeteksi plagiarisme saat ini menggunakan pertimbangan teks berbasis karakter yang canggih dan efisien. Sistem ini dapat mendeteksi kata per kata dan penyalinan teks dari sumber yang *original*. Akan tetapi, penataan ulang kalimat dan penggunaan makna tersirat yang kebanyakan ditemukan dalam penelitian menyebabkan perbandingan kesamaan karakter yang tidak cocok. Oleh karena itu, sistem pendeteksi plagiarisme tidak dapat mendeteksi bahwa dokumen tersebut plagiat atau tidak.

Untuk mengatasi masalah pada sistem pendeteksi saat ini, maka digunakan algoritma *decision tree*. Algoritma yang merupakan implementasi dari *decision tree* cukup beragam, Beberapa diantaranya yang paling sering digunakan saat ini antara lain; ID3, C4.5 dan CART. Tugas akhir ini menggunakan C4.5 yang merupakan implementasi dari algoritma *decision tree*. Alasan pemilihan C4.5 dikarenakan dapat menangani atau dapat mengatasi atribut-atribut yang bersifat kontinu dan bersifat diskrit, Algoritma C4.5 dapat mengatasi atribut *numeric*, data-data yang hilang serta data-data yang *error* dan algoritma ini dapat memangkas atau

menghapus cabang yang tidak diperlukan dan menggantinya dengan *node* daun. (Singh, Sonia dan Gupta,Priyanka, 2014)

Atas pertimbangan dari paragraf sebelumnya maka perlu adanya suatu sistem yang dapat melakukan pengujian deteksi plagiarisme yang mempertimbangkan probabilitas kemiripan kata dari segi struktur susunan SPO(Subjek,Predikat,Objek) dan dari segi makna yang tersirat dalam kalimat dengan menggunakan algoritma *decision tree* C4.5.

2 Rumusan Masalah

Berdasarkan penjelasan di atas, maka masalah yang diperoleh adalah sebagai berikut:

1. Kesamaan atau tidak pada makna kaliamt yang tersirat antara dokumen yang orisinal dengan dokumen yang plagiat.
2. Penataan ulang kalimat dengan struktur atau susunan yang berbeda pada plagiarisme.

3 Ruang Lingkup

Beberapa batasan dalam penerapan *system* deteksi plagiat dari struktur kata dan makna kalimat adalah sebagai berikut:

1. Sistem akan melakukan pengecekan terhadap dokumen per paragraf.
2. Sistem memproses dokumen yang berbahasa indonesia
3. Sistem mendeteksi dokumen secara *offline*.
4. Dokumen yang berformat *.pdf yang menjadi *file input* dalam pengecekan plagiat.

4 Tujuan Penelitian

Tujuan dari Penelitian *system* pengujian deteksi plagiat ini adalah

1. Sistem dapat melakukan pengujian deteksi plagiat dari struktur kata dan makna dengan menggunakan algoritma *Decision Tree* C4.5.

2. Mengeveluasi dan memperbaiki sistem yang ada saat ini yang hanya mengenali plagiat dari kata per kata.

5 Manfaat Penelitian

Manfaat dari penelitian dari sistem pengujian deteksi plagiat ini adalah:

1. Sistem yang digunakan dapat melakukan pengujian serta mendeteksi plagiat pada sebuah dokumen yang bermanfaat untuk penelitian kedepannya.
2. Dapat membantu masyarakat dibidang akademis seperti kampus-kampus khususnya agar mengetahui mahasiswa-mahasiswa yang melakukan plagiat terhadap tugas akhir (Skripsi, Tesis, disertasi) dan tugas-tugas kuliah lainnya
3. Sistem ini agar dapat dipergunakan dilingkungan akademis kampus STMIK-STIE MIKROSKIL.

6 Metodologi Pengembangan Perangkat Lunak

Langkah-langkah dalam pengerjaan tugas akhir ini, sebagai berikut:

1. Mengumpulkan dan mempelajari materi yang berhubungan dengan topik yang dibahas yaitu mengenai plagiarisme, *paraphrase* dan algoritma *decision tree*.
2. Pengembangan perangkat lunak dengan menggunakan model *waterfall* yang memiliki langkah kerja sebagai berikut:
    a. Memodelkan sistem yang akan dirancang dengan menggunakan alat bantu berupa UML (*Unified Modeling Language*).
    b. Merancang tampilan antarmuka pemakai (*user interface*) perangkat lunak.
    c. Membangun perangkat lunak dari permodelan sistem pada langkah kerja sebelumnya.
    d. Menguji perangkat lunak dan memperbaiki kesalahan yang muncul.
3. Menarik kesimpulan dari pengujian.
4. Menyusun laporan tugas akhir berdasarkan materi yang diperoleh dan perangkat lunak yang dibangun.

7 Tinjauan Pustaka

7.1 Plagiat

Plagiat atau plagiarisme adalah salah satu bentuk tindakan kecurangan, tetapi hal tersebut sulit untuk dijelaskan sehingga semua orang bahkan anak-anak bisa melakukannya tanpa menyadari dan memahami bahwa hal tersebut sebenarnya tidak dibenarkan. (Steven, Dowshen,2011)

Plagiat dapat dianggap sebagai tindakan melanggar hukum karena dikategorikan sebagai tindakan mencuri hak cipta seseorang. Di dunia pendidikan, para pelaku plagiat akan mendapatkan hukuman yang cukup berat bila pelaku plagiat adalah seorang mahasiswa atau seorang pendidik (Dosen, Guru) maka pelaku akan dikeluarkan dari universitas/sekolah tempat pelaku memperoleh ilmu atau yang berprofesi sebagai tenaga pendidik.

Berdasarkan pola, modus dan teknik plagiat dapat dikategorikan menjadi 4 jenis plagiarisme (Andreas Lako,2012) yaitu:

a) *Pertama*, Plagiarisme total adalah tindakan plagiat yang dilakukan oleh seorang penulis dengan cara melakukan penjiplakan hasil karya orang lain secara keseluruhan dan kemudian mengklaim bahwa hasil plagiat tersebut sebagai karyanya sendiri.
b) *Kedua*, Plagiarisme parsial adalah tindakan plagiat yang dilakukan oleh seorang penulis dengan cara melakukan penjiplakan hanya sebagian dari hasil karya orang lain, biasanya dalam melakukan plagiat seorang penulis hanya mengambil beberapa pernyataan, landasan teori, pembahasan sampai kesimpulan dan tanpa menyebutkan sumber aslinya.
c) *Ketiga*, Auto-Plagiasi (Self-plagiarisme) adalah tindakan plagiat yang dilakukan oleh seorang penulis terhadap hasil karyanya sendiri, baik hanya sebagian besar dari hasil karyanya atau secara keseluruhan.
d) *Keempat* ,Plagiarisme antarbahasa adalah tindakan plagiat yang dilakukan oleh seorang penulis dengan cara menerjemahkan suatu karya tulis yang berasal dari negara lain atau karya tulis yang berbahasa asing kemudian diterjemahkan ke dalam bahasa Indonesia.

7.2 Paraphrase

Paraphrase adalah strategi pemahaman makna suatu bentuk karya sastra dengan cara mengungkapkan kembali karya pengarang tertentu dengan menggunakan kata-kata yang berbeda dengan kata-kata yang digunakan pengarang. Tujuannya adalah menyederhanakan pemakaian kata atau kalimat (Aminuddin,2010:41)

Teknik *paraphrase* meliputi pengalihan bentuk dari hasil karya orang lain tapi tetap menyertakan sumber karya ilmiah.

a) Perubahan kata/frasa kunci dengan kata lain yang memiliki makna yang mirip atau sama. Proses ini menyangkut pemilihan kata yang memiliki persamaan arti(sinonim).
b) Perubahan bentuk kalimat asal dengan kalimat dengan susunan atau pola berbeda tanpa mengubah maknanya.
c) Perubahan bentuk wacana menjadi uraian yang lebih pendek berupa ringkasan atau rangkuman.

7.3 Decision Tree

Decision Tree atau pohon keputusan adalah salah satu algoritma yang digunakan untuk klasifikasi / seleksi. *Decision Tree* pada dasarnya adalah suatu sistem hirarki yang didekomposisi (penyederhanaan) *data training* dimana pada nilai atribut digunakan untuk membagi hirarki *data* pusat. Data pusat dilakukan secara rekursif didalam proses pengambilan keputusan, hingga *node leaf* (daun) terisi sesuai dengan jumlah *record*. *Decision Tree* akan membagi *node* sesuai kebutuhan variable dan memilih bagian yang paling *homogeny* dari sub node-node tersebut (Charu C. Aggarwal dan Chen Xiang Zhai, 2012)

Beberapa tahapan dalam proses decision tree (Tom M.Mitchell, McGraw Hill,1997):

a) Entropy

Untuk mendapatkan informasi yang tepat maka diperlukan pengukuran yang sering digunakan dalam klasifikasi yang disebut *entropy*. *Entropy* merupakan tingkat *parameter* untuk dapat mengetahui karakteristik dari impurty dan

homogeneity dari suatu kumpulan data. Dari hasil nilai *entropy* dapat dihitung nilai *information gain* dengan atribut *Boolean* sebagai berikut:

$$Entropy(S) = -P\oplus \log_2 P\oplus - P\ominus \log_2 P\ominus$$

Di mana:

S merupakan ruang data yang akan digunakan untuk *training*

$P\oplus$ merupakan jumlah yang mendukung pada *data sample*

$P\ominus$ merupakan jumlah yang tidak mendukung pada data sample untuk kriteria tertentu

Secara matematis, *entropy* dirumuskan dengan persamaan:

$$Entropy(S) = \sum_i^c - Pi \log_2 Pi$$

b) Information Gain

Setelah nilai *entropy* didapatkan di dalam suatu kumpulan *data training* maka dapat dilakukan pengukuran efektivitas suatu atribut dalam sebuah atribut dalam klasifikasi *data training*. Pengukuran yang digunakan disebut *information gain.* Secara matematis dapat dituliskan sebagai berikut:

$$Gain(S,A) = Entropy(S) - \sum_{vE} Value(A) \frac{|Sv|}{|S|} Entropy(Sv)$$

Dimana:

A: atribut

V: nilai untuk atribut

Values(A): himpunan pada atribut A

|Sv|: jumlah sampel untuk nilai V

|S|: jumlah seluruh sampel data

Entropy(Sv): entropy untuk sampel-sampel yang memiliki nilai v

c) Gain ratio

Modifikasi dari *information gain* untuk melakukan pengurangan data yang bias *Gain ratio* dihitung dengan menggunakan *split information* dengan persamaan

$$Split\ Information(S,A) = \sum_{i=1}^{c} - \frac{|Si|}{|S|} \log_2 \frac{|Si|}{|S|}$$

Gain ratio dirumuskan dengan persamaan

$$GainRatio(S,A) = \frac{Entropy(S) - \sum_{V \in Value(A)} \frac{|Sv|}{|S|} Entropy(Sv)}{SplitInformation(S,A)}$$

Dimana

S: himpunan sampel data

Si sampai Sc: sub bagian dari himpunan sampel data yang di bagi berdasarkan jumlah variasi nilai pada atribut *A*

7.4 C4.5

Algoritma *Decision Tree* C4.5 atau *Classification version* 4.5 adalah evolusi dari algoritma sebelumnya yaitu algoritma ID3, disebabkan karena kedua algoritma tersebut diciptakan oleh orang yang sama. (Quinlan,1983&1993). Algoritma C4.5 akan menghasilkan sebuah pohon keputusan(*Decision Tree*) terhadap data yang diberikan dan secara rekursif membagi data tersebut. Algoritma C4.5 dapat menentukan kemungkinan test dapat membagi data dan memilih data melalui tes yang memberikan *information gain* yang terbaik.(Singh, Sonia dan Gupta,Priyanka, 2014)

Kelebihan-kelebihan dari algoritma C4.5 dibandingkan dengan algoritma ID3 yaitu:

1. Algoritma C4.5 dapat menangani atau dapat mengatasi atribut-atribut yang bersifat kontinu dan bersifat diskrit.
2. Algoritma C4.5 dapat mengatasi atribut *numeric*, data-data yang hilang serta data-data yang *error*.
3. Algoritma dapat memangkas atau menghapus cabang yang tidak diperlukan dan menggantinya dengan *node* daun.

Secara umum ada 2 tahapan utama yaitu:

a) Information gain

Secara matematis:

$$Gain(S,A)=Entropy(S)-\sum_{i=1}^{n}\frac{|Si|}{|S|}*Entropy(Si)$$

Dengan:

S = Himpunan Kasus

A = Atribut

n = Jumlah partisi atribut A

|Si| = Jumlah kasus pada partisi ke-i

|S| = Jumlah kasus dalam S

b) Entropy

Secara matematis:

$$Entropy(S)=\sum_{i}^{c} - Pi \log_2 Pi$$

Dengan:

S = Himpunan Kasus

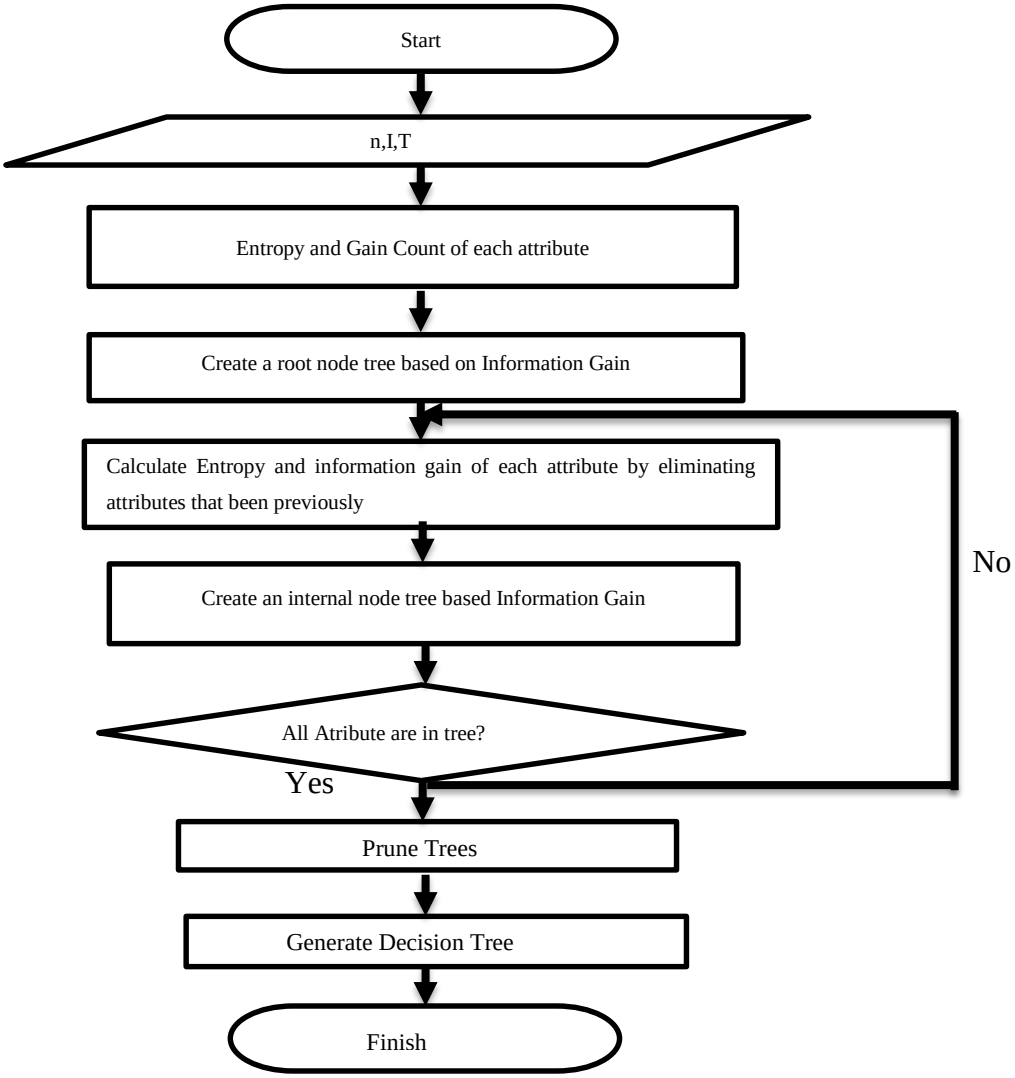
A = Fitur

n = Jumlah partisi S

pi = Proporsi dari Si terhadap S

Berikut ini adalah urutan proses kerja desain *flowchart* algoritma C4.5 (Setiawan,Edi dan Suhada,Siti, 2013)., antara lain :

1. Proses awal dimulai dari memasukkan *data training* atau atribut (n, T, I).
2. Menentukan nilai *entropy* dan nilai *gain* dari setiap atribut yang ada.
3. Membuat sebuah *root node* berdasarkan *information gain.*
4. Melakukan perhitungan nilai *entropy* dan *information gain*, dari hasil perhitungan tersebut atribut dengan gain tertinggi yang kemudian akan menjadi *root node* akan membentuk sebuah pohon*(Tree).*
5. Membuat sebuah *node* dalam Pohon berdasarkan nilai *information gain.*
6. Kemudian mengecek semua atribut telah ada di dalam pohon(*Tree),* jika atribut belum ada di pohon maka dilakukan proses kembali dari urutan 3-5.
7. Jika semua atribut telah ada di dalam pohon (*Tree*), maka dilakukan pengurangan *node tree* dari atribut yang ambigu dan bias dengan menggunakan persamaan *gain ratio Decision Tree.*
8. Kemudian dari proses yang dilakukan sebelumnya maka menghasilkan sebuah pohon keputusan*(Decision Tree).* Penjelasan yang lebih jelas dapat dilihat di gambar 7.4.1 Flowchart Algoritma *Decision Tree* C4.5.



Gambar. 7.4.1 Flowchart Algoritma *Decision Tree C4.5*

8 DAFTAR PUSTAKA

Aggarwal, Charu C. dan Zhai, Chen Xiang. (2012).*"ASurvey of Text Classification Algorithms"* (Online), *Chapter 6*. 52 halaman http://link.springer.com/chapter/10.1007%2F978-1-4614-3223-4_6 Diakses 15 February 2016

Juri D. Apresjan1, ed al. (2009)."*Semantic Paraprasing for Information Retrieval and Extraction*". http://link.springer.com/chapter/10.1007/978-3-642-04957-6_44

Lestarini, Ade Hapsari(2014)."Sederet Kasus Plagiarisme di Kampus", http://news.okezone.com/read/2014/02/25/373/946214/sederet-kasus-plagiarisme-di-kampus, Diakses 28 Maret 2016

McCabe Donald L. (2005). **"***Cheating among college and university students:A North American perspective***"**, *International Journal of Academic Integrity*.

Mitchell, Tom M.dan Hill ,McGraw .(1997)."*Machine Learning, Decision Tree Learning Chapter* 3"(Online) http://www.cs.princeton.edu/courses/archive/spr07/cos424/papers/mitchell-dectrees.pdf, http://link.springer.com/book/10.1007/978-1-4613-2279-5 diakses 26 Maret 2016

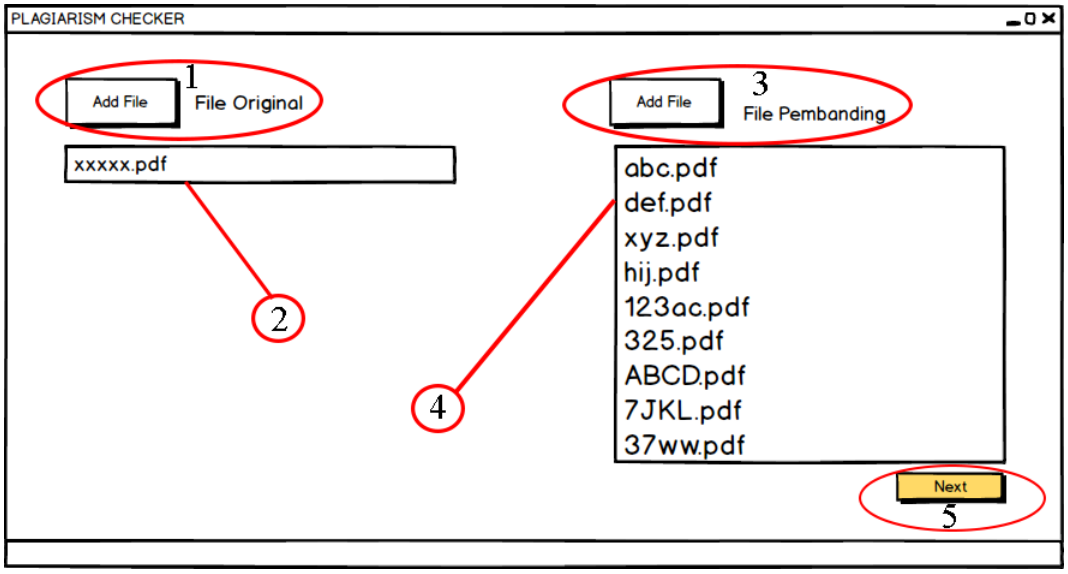
Ray, Sunly. (2015). "*Decision Tree Algorithms-Simplified*" http://www.analyticsvidhya.com/blog/2015/01/decision-tree-algorithms-simplified/ .Diakses 23 Maret 2016

Setiawan,Edi dan Suhada,Siti (2013) .*"Classification Needs Teachers Using Algorithm C4.5" Gorontalo State University*

Singh,Sonia dan Gupta,Priyanka(2014)."*Comparative Study ID3,CART and C4.5 Decision Tree Algorithm: A Survey*", *International Journal of Advanced Science and Technology(IJAIST)* Vol.27, No.27, July 2014.

9 Lampiran

Berikut ini adalah *Mockup* dari sistem deteksi tingkat plagiat:

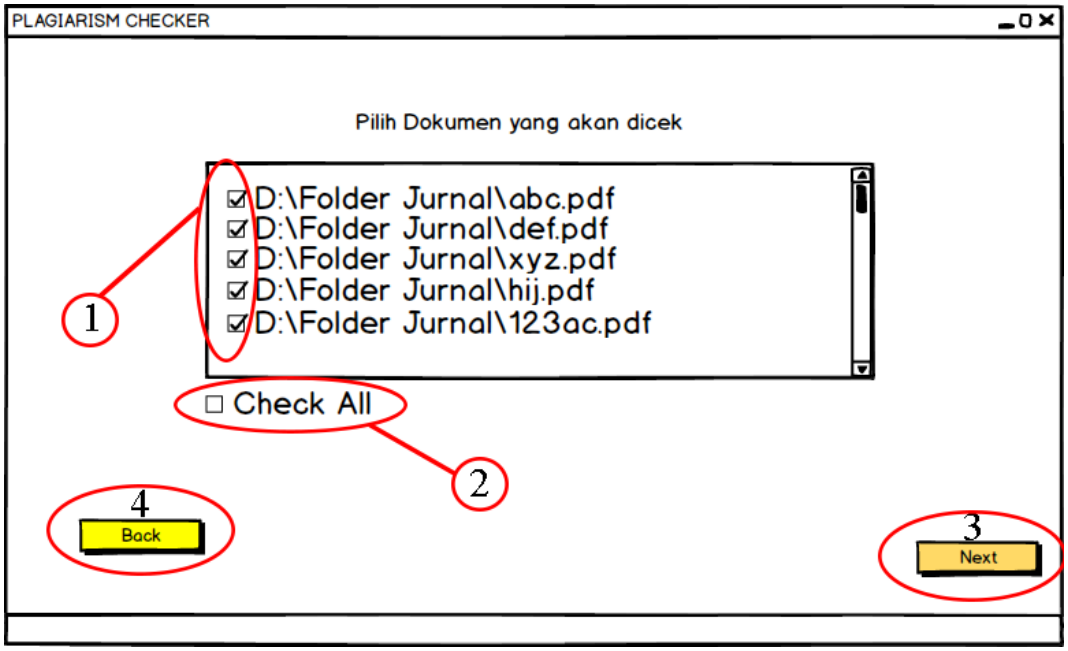


Gambar 9.1 form tampilan awal

Pada gambar 9.1 yang ada di atas merupakan form untuk melakukan pengambilan /memasukkan dokumen yang akan dilakukan pengecekan.

Keterangan:

1. Pada form tampilan awal tersebut merupakan tombol untuk melakukan pengambilan/dimasukkannya dokumen, dimana dokumen yang akan diambil adalah dokumen yang original dan jumlah dokumen yang dapat diambil hanya 1 dokumen.
2. Pada form tampilan awal tersebut merupakan textbox atau tempat dimana infomasi dari dokumen yang telah dimasukkan sebelumnya.
3. Pada form tampilan awal tersebut merupakan tombol yang memiliki fungsi yang sama dengan tombol pada no.1 yang membedakan adalah jumlah pengambilan/ penginputan dokumen yang lebih banyak dan dokumen yang dianggap plagiat.
4. Pada form tampilan awal tersebut merupakan textbox atau tempat dimana informasi dari setiap dokumen yang telah dimasukkan sebelumnya.
5. Pada form tampilan awal tersebut merupakan tombol untuk melanjutkan proses selanjutnya .

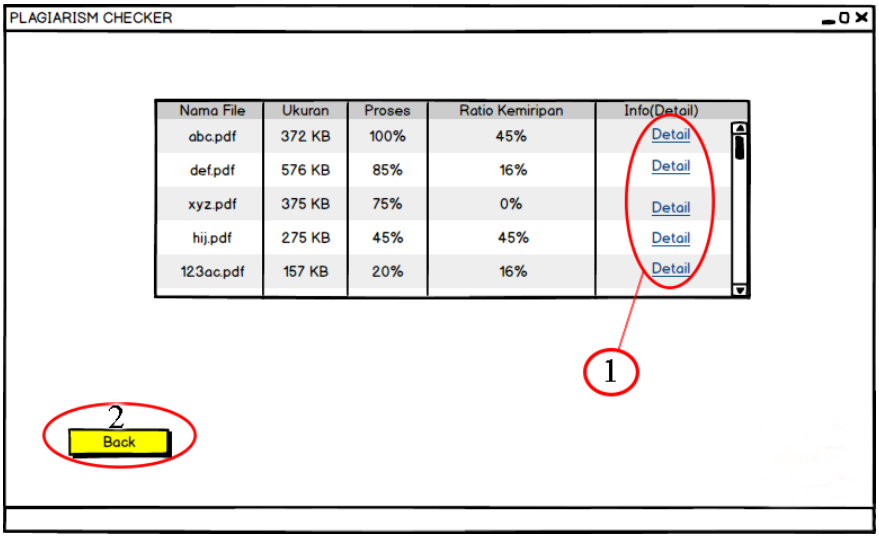


Gambar 9.2 form pemilihan dokumen

Pada gambar 9.2 yang ada diatas merupakan form untuk melakukan pemilihan dokumen yang akan dilakukan pengecekan.

Keterangan:

1. Pada form pemilihan dokumen merupakan *checkbox* yang digunakan untuk memilih beberapa dokumen yang diinginkan untuk dilakukan pengecekan.
2. Pada form pemilihan dokumen merupakan *checkbox* untuk memilih dokumen secara keseluruhan.
3. Pada form pemilihan dokumen merupakan tombol untuk melanjutkan proses selanjutnya ke form proses pengujian deteksi plagiat.
4. Pada form pemilihan dokumen merupakan tombol untuk kembali ke form tampilan awal.

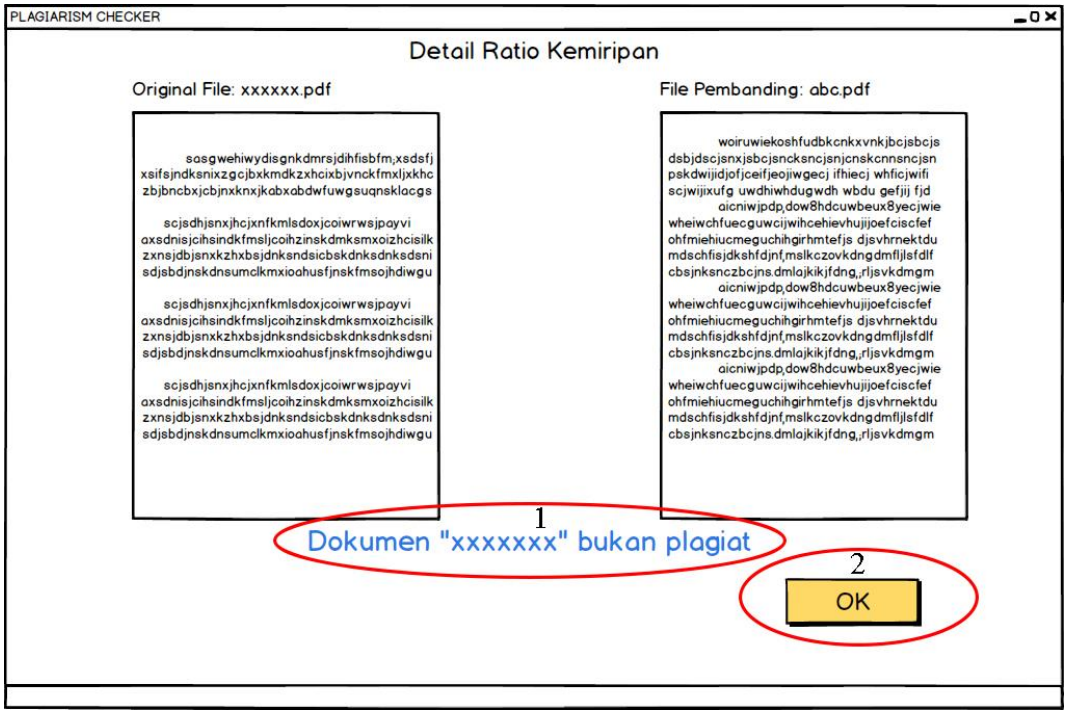


Gambar 9.3 form proses pengecekan deteksi plagiat

Pada gambar 9.3 merupakan form pengujian terhadap dokumen-dokumen yang sebelumnya telah dipilih serta form ini menampilkan persentase kemiripan dokumen pembanding / dianggap plagiat dengan dengan dokumen orisinal

Keterangan:

1. Pada form proses merupakan *linklabel* yang digunakan untuk melihat info lebih lanjut dimana info mengenai dokumen tersebut dikategorikan plagiat atau tidak.
2. pada form proses merupakan tombol untuk kembali ke form sebelumnya yaitu form pemilihan dokumen.

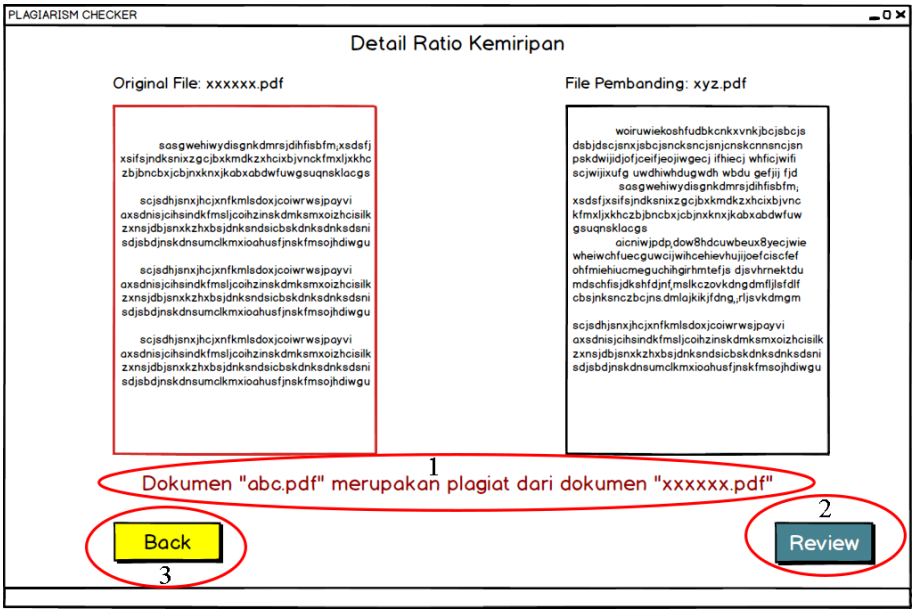


Gambar 9.4 form detail rasio kemiripan

Pada gambar 9.4 merupakan form informasi bahwa dokumen tersebut bukan plagiat

Keterangan:

1. pada form rasio kemiripan tersebut merupakan label yang memberikan informasi bahwa dokumen tersebut bukan plagiat dengan tanda tulisan berwarna biru.
2. pada form rasio kemiripan tersebut merupakan tombol untuk kembali ke form sebelumnya yaitu form proses pengecekan dokumen.

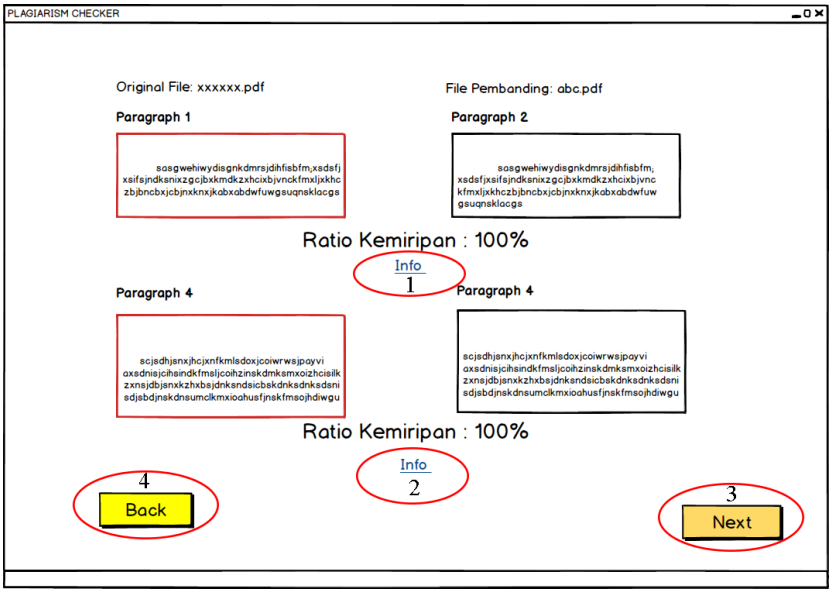


Gambar 9.5 form rasio kemiripan

Pada gambar 9.5 merupakan form yang memberikan informasi jika suatu dokumen dikategorikan sebagai plagiat

Keterangan :

1. pada form rasio kemiripan tersebut merupakan label yang memberikan informasi bahwa dokumen tersebut dikategorikan sebagai plagiat.
2. pada form rasio kemiripan tersebut merupakan tombol untuk melanjutkan ke form selanjutnya yaitu form *review* dimana yang direview adalah paragraf.
3. pada form rasio kemiripan tersebut merupakan tombol untuk kembali ke form sebelumnya yaitu form proses pengecekan dokumen.



Gambar 9.6 *review* dokumen

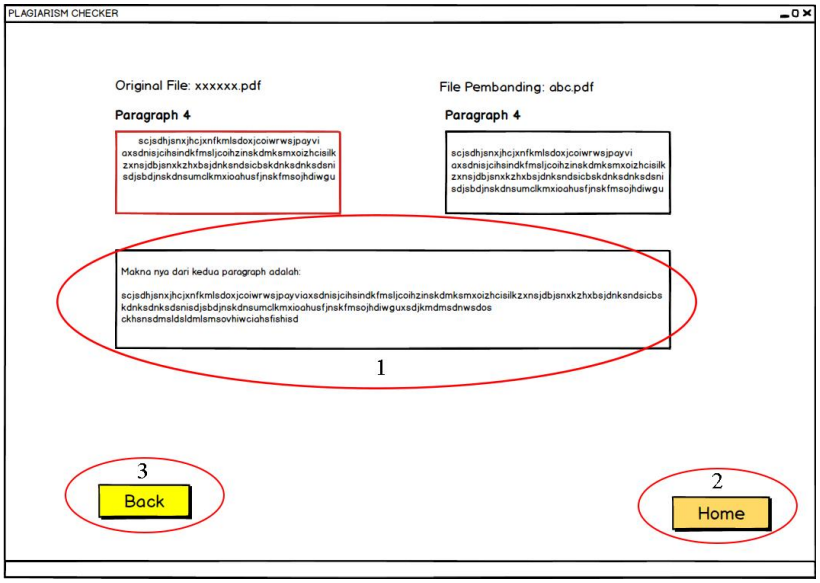
Pada gambar 9.6 merupakan form yang memberikan informasi tentang suatu paragraf dari suatu dokumen yang mana memiliki kemiripan susunan kata dengan paragraf di dokumen yang lain dalam artian dokumen yang original

Keterangan :

1, 2 pada form *review* dokumen tersebut merupakan *linklabel* dimana digunakan untuk melihat informasi selanjutnya mengenai makna dari suatu paragraf.

3 pada form *review* dokumen tersebut merupakan tombol untuk ke form awal.

4 pada form *review* dokumen tersebut merupakan tombol untuk kembali ke form detail rasio kemiripan.



Gambar 9.7 form *info*

Pada gambar 9.7 merupakan form yang menampilkan informasi mengenai paragraf yang mana yang memiliki susunan kata yang sama dan memiliki makna yang sama juga

Keterangan:

1. pada form *info* tersebut merupakan sebuah *textbox* yang menampilkan maksud dan makna dari paragraph.
2. pada form *info* tersebut merupakan tombol untuk ke form awal .
3. pada form *info* tersebut merupakan tombol untuk kembali ke form sebelumnya yaitu form *review* dokumen.