BAB 1
PENDAHULUAN

1.1. Latar Belakang

Di era digital sekarang ini, telah terjadi lonjakan produksi data digital yang sangat tinggi. Berdasarkan data dari laman forbes.com dalam survei pada tahun 2018, terdapat 2.5 quintillion bytes data yang diproduksi setiap hari dan produksi data digital tersebut akan terus bertambah setiap tahun (Marr, 2018). Dari quintillion data tersebut, sebagian besar merupakan data berformat PDF. Berdasarkan data dari PDF Association pada April 2016, telah diproduksi: (Association, 2018)

- 2,2 miliar file PDF di web
- 20 miliar file PDF di Dropbox
- Airbus, Boeing, dan Departemen Kehakiman AS masing masing memiliki lebih dari 1 miliar PDF
- 2 miliar PDF dibuka setiap tahun di outlook.com
- 73 juta file PDF baru disimpan setiap hari di Google Drive dan Email
- 60% lampiran non-gambar adalah PDF di Outlook Exchange Enterprise.

Hal ini menunjukkan bahwa format *file* PDF sangat penting dan sangat dibutuhkan. Alasan yang paling penting mengapa banyak orang

menggunakan PDF dibandingkan dengan format lain adalah memiliki fleksibilitas yang tinggi. Salah satu pemanfaatan PDF adalah digunakan sebagai format *file* pada tugas mahasiswa. Dengan menggunakan PDF, *file* tersebut tidak akan berubah formatnya walapun di buka dari perangkat yang berbeda.

Dizaman pandemi COVID-19 sekarang ini, banyak kampus yang menyelenggarakan proses belajar mengajar online atau daring menggunkan platform Google Classroom. Tugas mahasiswa berformat PDF akan dikumpulkan pada platform Google Classroom. Namun setelah tugas mahasiswa dikumpulkan dan diperiksa oleh dosen, ternyata banyak tugas yang plagiasi. Dimana, jika tugas mahasiswa tersebut diperiksa secara manual akan membutuhkan waktu yang lama dan melelahkan. Maka dibutuhkan sebuah sistem yang dapat menyaring tugas mahasiswa yang plagiasi dan tidak plagiasi. Sehingga dosen hanya perlu memeriksa tugas siswa yang plagiasi saja.

Teknik yang biasanya digunakan untuk proses penyaringan adalah teknik Approximate Matching. Approximate Matching merupakan sebuah teknik yang digunakan untuk menemukan kesamaan di antara beberapa file berdasarkan skor kesamaan. Beberapa algoritme yang biasa diguanakan adalah Sd-Hash dan Ssdeep. Namun algoitme ini telah terbukti rentan terhadap serangan aktif (Chang, et al., 2015) (Roussev,

2011). Sehingga dibutuhkan algoritme baru yang aman terhadap serangan aktif dan memiliki akurasi yang tinggi untuk menghitung nilai kesamaan diantara beberapa *file*.

Algoritme Frequency Based Hashing-S atau FBHash merupakan algortime baru yang dipublikasikan pada tahun 2019. Pada algoritme FBHash, setiap *byte* dari *chunk* berkontribusi pada skor akhir dan pengaruhnya terhadap skor akhir tergantung pada pentingnya suatu dokumen (Chang, et al., 2019). Untuk membuat skor kesamaan benar benar rendah atau mendekati nol hampir setiap *chunk* harus di modifikasi. Secara keamanan, algoritme FbHash lebih baik dari algoritme SdHash dan Ssdeep, oleh sebab itu penulis memilih untuk menggunakan algoritme FbHash ini.

Algoritme FbHash dibagi menjadi dua versi, yaitu versi FbHash-B untuk mengukur kesamaan ditingkat *byte* seperti format text dan FbHash-S untuk mengukur *syntactic matching* dan menggunakan informasi internal dokumen untuk mengukur kesamaan. Algoritme ini digunakan untuk format *file* terkompresi seperti PDF. Karena object penelitian dari penulis adalah *file* yang terkompresi dalam format PDF, maka penulis menggunakan algoritme FbHash-S.

Berdasarkan penelitian sebelumnya dari Chang, et al., 2019 yang berjudul “FbHash: A New Similarity Hashing Scheme for Digital

Forensics" menyatakan bahwa skema FbHash aman terhadap serangan aktif dan mendeteksi kesamaan dengan akurasi 98% dan memiliki tingkat akurasi 50% lebihi tinggi dari skema lain untuk format data yang terkompresi. Berdasarkan masalah-masalah yang telah diuraikan diatas maka peneliti mencoba melakukan penelitian yang diberi judul : **SISTEM DETEKSI PLAGIASI MENGGUNAKAN ALGORITME FREQUENCY BASED HASHING-S PADA FILE PDF.**

1.2. Identifikasi Masalah

Adapun identifikasi masalah sebagai berikut :

1. Memeriksa tugas mahasiswa secara manual memerlukan waktu yang lama dan melelahkan
2. Algoritme Approximate Matching yang ada sekarang ini rentan terhadap serangan aktif
3. Dibutuhkan algoritme baru yang aman terhadap serangan aktif dan memiliki tingkat akurasi yang tinggi

1.3. Batasan Masalah

Untuk membatasi ruang lingkup penelitian, maka dirumuskan Batasan masalah sebagai berikut:

1. Data yang digunakan adalah file PDF yang didalamnya hanya terdapat teks dan angka, tidak mengandung tabel, diagram

dan gambar ataupun rumus matematika. Hal ini dilakukan untuk memudahkan dalam mengidentifikasi teks pada *file* tersebut.

2. Penelitian ini hanya akan menampilkan nama data beserta tingkat plagiasinya, bukan melakukan penyaringan secara langsung dengan membedakan *file* yang plagiasi dan tidak plagiasi.

1.4. Rumusan Masalah

Berdasarkan latar belakang di atas, maka dirumuskan masalah sebagai berikut :

1. Bagaimana perancangan system deteksi plagiagiasi menggunakan algoritma frequency based hashing-S pada file PDF?
2. Bagaimanakah implementasi algoritme Frequency Based Hashing-S pada sistem deteksi plagiasi untuk menghitung nilai skor kesamaan tugas mahasiswa berformat PDF?
3. Bagaimanakah akurasi perhitungan nilai skor kesamaan dengan mengimplementasikan algoritme Frequency Based Hashing-S pada file PDF?
4. Bagaimanakah integritas data *file* PDF dengan mengimplementasikan algoritme Frequency Based Hashing-S

pada sistem deteksi plagiasi?

1.5. Tujuan Penelitian

Manfaat dari penelitian ini diharapkan dapat menjadi dasar untuk pemeriksaan dokumen dengan format PDF yang membutuhkan metode penyortiran untuk memilah dokumen yang akan diperiksa secara manual sesuai dengan nilai skor kesamaan atau skor plagiasinya. Sehingga dokumen dengan tingkat plagiasi tinggi atau nilai skor kesamaan tinggi tidak perlu diperiksa lagi untuk mengurangi waktu pemeriksaan dokumen secara manual. Dokumen dalam jumlah banyak dapat disaring menjadi dokumen dalam jumlah kecil yang paling bermanfaat tanpa mengurangi tingkat akurasi kesamaan dan dapat menjamin keamanan data.

1.6. Manfaat Penelitian

Manfaat dari penelitian ini diharapkan dapat menjadi dasar untuk pemeriksaan dokumen dengan format PDF yang membutuhkan metode penyortiran untuk memilah dokumen yang akan diperiksa secara manual sesuai dengan nilai skor kesamaan atau skor plagiasinya. Sehingga dokumen dengan tingkat plagiasi tinggi atau nilai skor kesamaan tinggi tidak perlu diperiksa lagi untuk mengurangi waktu pemeriksaan dokumen secara manual. Dokumen dalam jumlah banyak dapat disaring menjadi dokumen dalam jumlah kecil yang

paling bermanfaat tanpa mengurangi tingkat akurasi kesamaan dan dapat menjamin keamanan data.