

**IMPLEMENTASI ALGORITMA *RANDOM FORESTS* UNTUK
KLASIFIKASI *SPAM* PADA CITRA DAN *TEXT*
INSTAGRAM®**

TUGAS AKHIR



RIZKY NOVRIYEDI PUTRA

1132001001

**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN ILMU KOMPUTER
UNIVERSITAS BAKRIE
JAKARTA
2017**

**IMPLEMENTASI ALGORITMA *RANDOM FORESTS* UNTUK
KLASIFIKASI *SPAM* PADA CITRA DAN *TEXT*
INSTAGRAM®**

TUGAS AKHIR

**Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana
Komputer**



RIZKY NOVRIYEDI PUTRA

1132001001

**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN ILMU KOMPUTER
UNIVERSITAS BAKRIE
JAKARTA**

2017

HALAMAN PERNYATAAN ORISINALITAS

**Tugas Akhir ini adalah hasil karya saya sendiri,
dan semua sumber baik yang dikutip maupun dirujuk
telah saya nyatakan dengan benar.**

Nama : Rizky Novriyedi Putra

NIM : 1132001001

Tanda Tangan :



Tanggal : 12 September 2017

HALAMAN PENGESAHAN


Tugas Akhir ini diajukan oleh:

Nama : Rizky Novriyedi Putra
NIM : 1132001001
Program Studi : Informatika
Fakultas : Teknik dan Ilmu Komputer
Judul Skripsi : Implementasi Algoritma *Random Forests* Untuk Klasifikasi
Spam Pada Citra dan *Text* Instagram

Telah berhasil dipertahankan di hadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana Komputer pada Program Studi Informatika Fakultas Teknik dan Ilmu Komputer Universitas Bakrie

DEWAN PENGUJI

Pembimbing : Guson P. Kuntarto, ST, Msc.

()

Penguji I : Berkah I. Santoso, S.T, M.T.I.

()

Penguji II : Prof. Dr. Hoga Saragih, ST, MT.

()

Ditetapkan di : Jakarta

Tanggal : 12 September 2017

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI

Sebagai sivitas akademik Universitas Bakrie, saya yang bertanda tangan di bawah ini:

Nama : Rizky Novriyedi Putra
NIM : 1132001001
Program Studi : Informatika
Fakultas : Teknik dan Ilmu Komputer

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Bakrie **Hak Bebas Royalti Noneksklusif (*Non-exclusive Royalty-Free Right*)** atas karya ilmiah saya yang berjudul:

Implementasi Algoritma *Random Forests* Untuk Klasifikasi *Spam* Citra dan *Text* Instagram

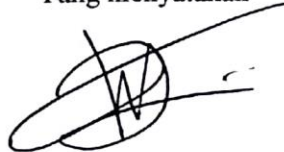
beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Universitas Bakrie berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta untuk kepentingan akademis.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Jakarta

Pada tanggal : 12 September 2017

Yang menyatakan



(Rizky Novriyedi Putra)

**IMPLEMENTASI ALGORITMA *RANDOM FORESTS* UNTUK
KLASIFIKASI *SPAM* PADA CITRA DAN *TEXT* INSTAGRAM**

Rizky Novriyedi Putra

ABSTRAK

Pada saat sekarang ini media sosial memegang peranan penting dalam berkomunikasi dan berbagi pengalaman, Instagram merupakan sebuah *platform* media sosial yang menggunakan gambar, dan *text* dalam komunikasi sesama pengguna. Informasi yang tersebar di Instagram dapat berupa promosi barang, dan konten yang tidak berhubungan dengan penerima dan lain sebagainya, hal tersebut merupakan *spam* yang dapat mengganggu kenyamanan. Permasalahan tersebut dapat dicegah dengan mengklasifikasi data gambar dan *text* pada Instagram kedalam kategori *spam* dan *non-spam*. Untuk melakukan hal tersebut diperlukannya sebuah *dataset* yang dapat diproses menggunakan metode *gray level coocurance matrix* (GLCM) untuk data gambar dan *term frequency invers document frequency* (TF/IDF) untuk data *text*. Proses klasifikasi pada penelitian ini menggunakan 250 data *spam* dan 250 data *non-spam*, dimana masing-masing *dataset images* dan *dataset text* berjumlah 500 data. Algoritma *random forests* digunakan sebagai metode klasifikasi dalam *machine learning* yang memiliki tahapan *random feature selection*, dan *bootstrap aggregation* dalam pembentukan model, untuk proses klasifikasi *random forests* menggunakan *majority vote*, hasil penelitian ini memiliki *f-measure* rata-rata 70% untuk *dataset* gambar dan rata-rata 60% untuk *dataset text* dengan melakukan perubahan parameter *bootstrap aggregation* menjadi 3 diantaranya 1/2, 1/6, 2/3 data dari *dataset*.

Kata Kunci: Instagram, *Gray Level Coocurance Matrix* (GLCM), *Term Frequency/ Invers Document Frequency* (TF/IDF), *Machine Learning*, *Random Forests*, *Random Feature Selection*, *Bagging*, *Majority Vote*.

**IMPLEMENTATION OF RANDOM FORESTS ALGORITHM FOR SPAM
CLASSIFICATION ON IMAGE AND TEXT INSTAGRAM**

Rizky Novriyedi Putra

ABSTRACT

Today, social media is an important part of communication and sharing experience to the people in network. One of that is Instagram, a social media platform that using picture and text in communication to each other user. Information spread on the Instagram have a many variation like a promotion, and content that doesn't related to other user and etc, that will disturb the comfort to some user. The problems can be solved with classification the data of text and picture to spam and non-spam category. To do that, we'll need some dataset that can be processed with method of Gray Level Concurrence Matrix (GLCM) to data picture and Term Frequency Invers Document Frequency (TF/IDF) to data text. The process classification on this study using 250 spam and 250 non-spam data where in each other data set total is 500 data. Algorithm Random Forest used in classification method for machine learning that have a random feature selection and bootstrap aggregation in forming the model, for process classification random forest used majority vote. The result of this study have Average F-Measure 70% on picture dataset and 60% on text dataset with 3 (three) parameter bootstrap aggregation among others, 1/2, 1/6, 2,3 data from dataset.

Key Words: Instagram, Gray Level Coocurance Matrix (GLCM), Term Frequency/ Invers Document Frequency (TF/IDF), Machine Learning, Random Forests, Random Feature Selection, Bagging, Majority Vote.

DAFTAR ISI

HALAMAN JUDUL	ii
HALAMAN PERNYATAAN ORISINALITAS	Error! Bookmark not defined.
HALAMAN PENGESAHAN	Error! Bookmark not defined.
UNGKAPAN TERIMA KASIH	iv
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI	Error! Bookmark not defined.
ABSTRAK.....	vi
ABSTRACT.....	vii
DAFTAR ISI.....	viii
DAFTAR GAMBAR	xi
DAFTAR TABEL.....	xiii
DAFTAR RUMUS	xiv
DAFTAR LAMPIRAN.....	xv
BAB I PENDAHULUAN.....	1
1.1. Latar Belakang	1
1.2. Rumusan Masalah.....	4
1.3. Tujuan Penelitian	4
1.4. Ruang Lingkup Penelitian.....	4
1.5. Sistematika Penulisan	5
BAB II TINJAUAN PUSTAKA	6
2.1. Penelitian Terkait	6
2.2. Spam	10
2.3. Feature Extraction.....	11
2.4. Machine Learning	18
BAB III METODOLOGI PENELITIAN	25

3.1. Studi Literatur	27
3.2. Merumuskan Tujuan Penelitian	27
3.3. Pembentukan Dataset.....	28
3.4. Implementasi Random Forests.....	33
3.5. Hasil Implementasi dan Pembahasan.....	35
3.6. Penyusunan Laporan Hasil Penelitian	35
BAB IV IMPLEMENTASI DAN HASIL PENELITIAN.....	36
4.1. Pengumpulan Data	36
4.2. Ekstraksi Fitur	39
4.2.1. Ekstraksi Fitur Images	39
4.2.2. Ekstraksi Fitur Text.....	40
4.3. Labeling	41
4.4. Dataset.....	42
4.4.1. Training Data	42
4.4.2. Testing Data	43
4.5. Pembentukan Random Forests.....	43
4.5.1. Random Feature Selection	44
4.5.2. Bootstrap Aggregation (Bagging).....	44
4.5.3. Pembentukan Pohon.....	45
4.6. Eksperimen	46
4.6.1. Hasil Model Training	47
4.6.2. Hasil Uji Coba Data Images	48
4.6.3. Hasil Uji Coba Data Text.....	52
4.7. Pembahasan.....	55
BAB V SIMPULAN DAN SARAN.....	56
5.1. Simpulan	56

5.2 Saran	57
DAFTAR PUSTAKA	58
LAMPIRAN.....	62

DAFTAR GAMBAR

Gambar 2. 1 Proses Kuantisasi [10].....	13
Gambar 2. 2 Empat Sudut GLCM [22].....	14
Gambar 2. 3 Proses Pembentukan <i>Co-occurrence Matrix</i> [10].....	14
Gambar 2. 4 Proses Pembentukan Matrik Simetris [10].....	15
Gambar 2. 5 Proses Normalisasi Matrik Simetris [10].....	15
Gambar 2. 6 <i>Flowchart</i> Algoritma <i>Random Forests</i> [30].....	21
Gambar 2. 7 <i>Decision tree</i> [37].....	22
Gambar 3. 1. Tahapan Penelitian	25
Gambar 3. 2 Pembentukan <i>Dataset</i> dan <i>Implementasi Random Forests</i>	26
Gambar 3. 3. Hasil <i>Crawling</i> [34].	30
Gambar 3. 4. Tahapan <i>Text Preprocessing</i>	31
Gambar 3. 5. Proses Klasifikasi Algoritma <i>Random Forests</i>	34
Gambar 3. 6. Tabel <i>Confusion Matrix Images</i> dan <i>Text</i>	35
Gambar 4. 1 Potongan JSON hasil <i>Crawling</i>	36
Gambar 4. 2. Potongan <i>description</i>	37
Gambar 4. 3. Potongan <i>Raw Data Images Non-Spam</i>	37
Gambar 4. 4. Potongan <i>Raw Data Images Spam</i>	38
Gambar 4. 5. Potongan <i>Raw Data Text Non-Spam</i>	38
Gambar 4. 6. Potongan <i>Raw Data Text Spam</i>	38
Gambar 4. 7. Potongan Hasil <i>Preprocess Class GLCM Data Images Non-Spam</i> ..	39
Gambar 4. 8. Potongn Hasil <i>Preprocess Class GLCM Data Images Spam</i>	39
Gambar 4. 9. Parameter <i>Query</i>	40
Gambar 4. 10. Kumpulan <i>File JSON</i> hasil <i>Crawling</i>	40

Gambar 4. 11. Potongan Hasil <i>Preprocess Class</i> TF-IDF.....	41
Gambar 4. 12. <i>Script Labeling Data</i>	42
Gambar 4. 13. <i>Pseudocode</i> Algoritma <i>Random Forests</i> [36].....	43
Gambar 4. 14. Implementasi Algoritma <i>Random Forests</i>	44
Gambar 4. 15. Contoh Model <i>Random Forests Images</i>	45
Gambar 4. 16. Contoh Model <i>Random Forests Text</i>	46
Gambar 4. 17. Potongan Data Hasil Uji Coba	47
Gambar 4. 18. Hasil Model Data <i>Text</i>	47
Gambar 4. 19. Hasil Model Data <i>Images</i>	48
Gambar 4. 20. Hasil Rata-Rata Uji Coba $2/3$ <i>Bagging Images</i>	49
Gambar 4. 21. Hasil Rata-Rata Uji Coba $1/2$ <i>Bagging Images</i>	50
Gambar 4. 22. Hasil Rata-Rata Uji Coba $1/6$ <i>Bagging Images</i>	51
Gambar 4. 23. Hasil Rata-Rata Uji Coba $2/3$ <i>Bagging Text</i>	52
Gambar 4. 24. Hasil Rata-Rata Uji Coba $1/2$ <i>Bagging Text</i>	53
Gambar 4. 25. Hasil Rata-Rata Uji Coba $1/6$ <i>Bagging Text</i>	54

DAFTAR TABEL

Tabel 2. 1 Rangkuman Hasil Penelitian Terkait	9
Tabel 2. 2 Pengelompokan Warna [10].	13
Tabel 2. 3 <i>Confusion Matrix</i> 2x2 [12].	23

DAFTAR RUMUS

Persamaan (2.1) *Contrats*

Persamaan (2.2) *Correlation*

Persamaan (2.3) *Energy (Angular Second Moment)*

Persamaan (2.4) *Entropy*

Persamaan (2.5) *Homogeneity*

Persamaan (2.6) TF (*Term Frequency*)

Persamaan (2.7) IDF (*Inverse Document Frequency*)

Persamaan (2.8) TF-IDF

Persamaan (2.9) *Information Gain*

Persamaan (2.10) *Gain*

Persamaan (2.11) *Entropi*

Persamaan (2.12) *F-Measure*

Persamaan (2.13) *Recall*

Persamaan (2.14) *Precision*

DAFTAR LAMPIRAN

- Lampiran 1: *Raw Data Non-Spam Images*
- Lampiran 2: *Raw Data Spam Images*
- Lampiran 3: *Raw Data Spam Text*
- Lampiran 4: *Raw Data Non-Spam Text*
- Lampiran 5: *Dataset Images*
- Lampiran 6: *Dataset Text*
- Lampiran 7: *Hasil Implementasi Random Forests Data Training Images*
- Lampiran 8: *Hasil Implementasi Random Forests Data Training Text*
- Lampiran 9: *Hasil Implementasi Random Forests Data Testing Images*
- Lampiran 10: *Hasil Implementasi Random Forests Data Testing Text*
- Lampiran 11: *Source Code Class GLCM*
- Lampiran 12: *Source Code Class TF-IDF*
- Lampiran 13: *Source Code Class Random Forests*
- Lampiran 14: *Source Code Class Confusion Matrix*