

**DETEKSI ULASAN PALSU PADA TOKOPEDIA DAN
SHOPEE MENGGUNAKAN *NATURAL LANGUAGE
PROCESSING***

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana Program
Studi Informatika



Disusun oleh:

RAHMAT MUFLIH ARIFIN

18.11.2242

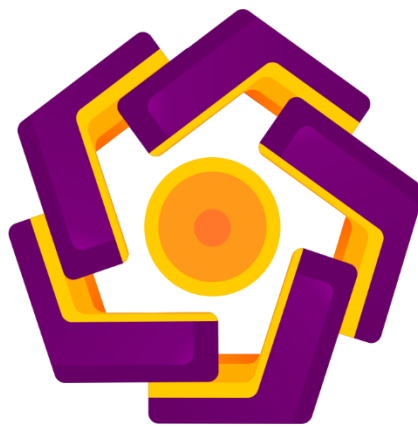
Kepada

**PROGRAM SARJANA
PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2023**

**DETEKSI ULASAN PALSU PADA TOKOPEDIA DAN
SHOPEE MENGGUNAKAN *NATURAL LANGUAGE
PROCESSING***

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



Disusun oleh:

**RAHMAT MUFLIH ARIFIN
18.11.2242**

**PROGRAM SARJANA
PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2023**

HALAMAN PERSETUJUAN

SKRIPSI

**DETEKSI ULASAN PALSU PADA TOKOPEDIA DAN
SHOPEE MENGGUNAKAN *NATURAL LANGUAGE
PROCESSING***

yang dipersiapkan dan disusun oleh

Rahmat Muflih

18.11.2242

Telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 20 Januari 2023

Dosen Pembimbing,

Theopilus Bayu Sasongko, S.Kom, M.Eng

NIK. 190302375

HALAMAN PENGESAHAN

HALAMAN PENGESAHAN

SKRIPSI

**DETEKSI ULASAN PALSU PADA TOKOPEDIA DAN
SHOPEE MENGGUNAKAN *NATURAL LANGUAGE
PROCESSING***

yang dipersiapkan dan disusun oleh

Rahmat Muflih Arifin

18.11.2242

Telah dipertahankan di depan Dewan Penguji
pada tanggal 20 Januari 2023

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

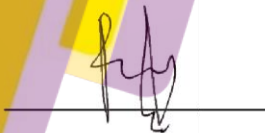
Andi Sunyoto, M.Kom., Dr.
NIK. 190302052



Ika Nur Fajri, M.Kom.
NIK. 190302268



Theopilus Bayu Sasongko, S.Kom, M.Eng
NIK. 190302375



Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 20 Januari 2023

DEKAN FAKULTAS ILMU KOMPUTER

Hanif Al Fatta, M.Kom.
NIK. 19030209



HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : **Rahmat Muflih Arifin**
NIM : **18.11.2242**

Menyatakan bahwa Skripsi dengan judul berikut:

Deteksi Ulasan Palsu Pada Tokopedia dan Shopee Menggunakan *Natural Language Processing*

Dosen Pembimbing : Theopilus Bayu Sasongko, S.Kom, M.Eng

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 20 Februari 2023

Yang Menyatakan,

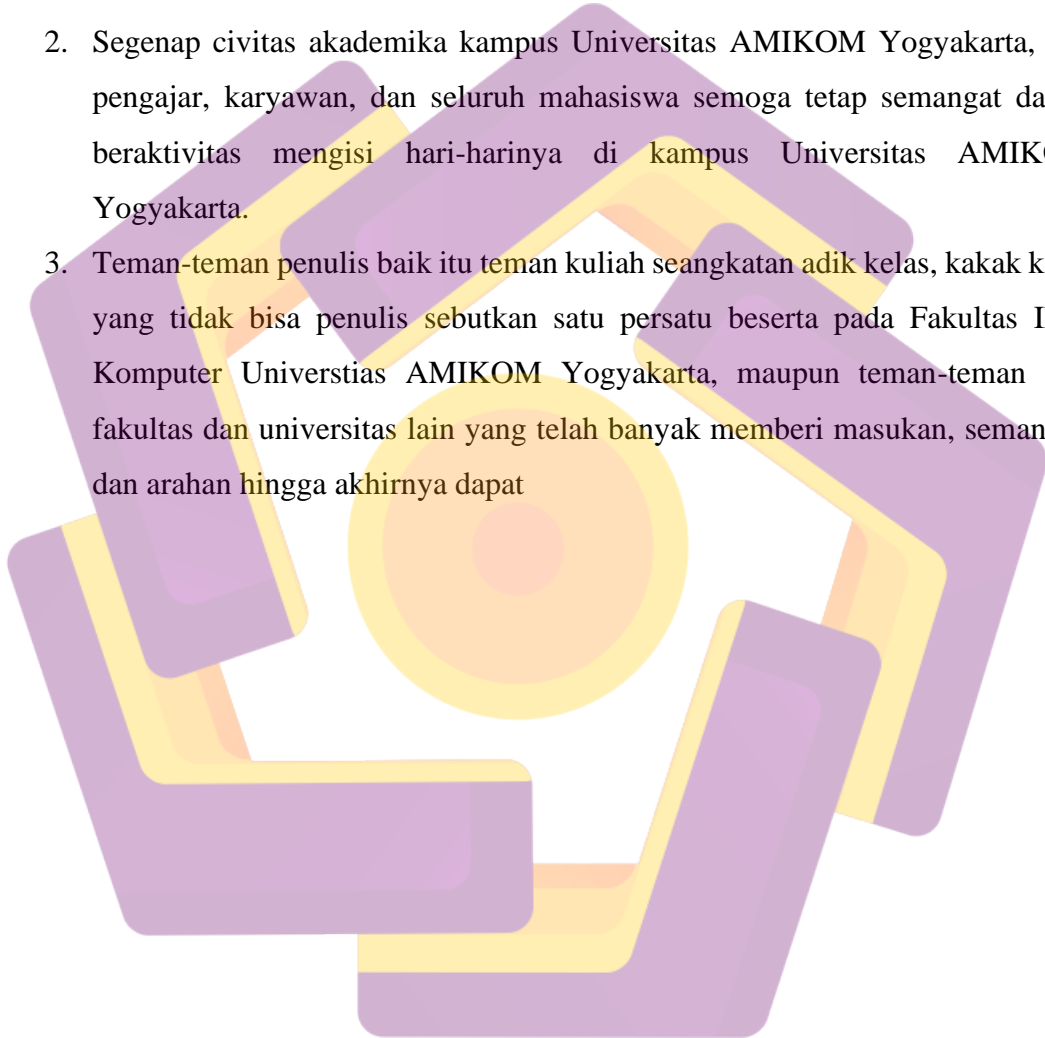


Rahmat Muflih Arifin

HALAMAN PERSEMBAHAN

Dengan rasa syukur yang mendalam, dengan telah diselesaikannya Skripsi ini Penulis mempersembahkannya kepada:

1. Keluarga besar penulis yang telah senantiasa membantu menyelesaikan skripsi ini.
2. Segenap civitas akademika kampus Universitas AMIKOM Yogyakarta, staf pengajar, karyawan, dan seluruh mahasiswa semoga tetap semangat dalam beraktivitas mengisi hari-harinya di kampus Universitas AMIKOM Yogyakarta.
3. Teman-teman penulis baik itu teman kuliah seangkatan adik kelas, kakak kelas yang tidak bisa penulis sebutkan satu persatu beserta pada Fakultas Ilmu Komputer Universtias AMIKOM Yogyakarta, maupun teman-teman dari fakultas dan universitas lain yang telah banyak memberi masukan, semangat, dan arahan hingga akhirnya dapat



KATA PENGANTAR

Assalamu'alaikum Wr. Wb.

Alhamdulillahil'alaamiin, puji syukur senantiasa penulis panjatkan atas kehadiran Allah SWT yang telah melimpahkan rahmat dan hidayahNya, sehingga penulis dapat menyelesaikan tugas akhir yang berjudul “Deteksi Ulasan Palsu Pada Tokopedia dan Shopee Menggunakan *Natural Language Processing*” sebagai salah satu persyaratan yang harus dipenuhi dalam menyelesaikan jenjang Strata Satu atau S1 di Jurusan Informatika, Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta.

Shalawat serta salam InsyaAllah selalu tercurah kepada Nabi Muhammad SAW serta para sahabat dan pengikutnya sampai akhir zaman. Penyelesaian tugas akhir ini tidak terlepas dari bantuan, arahan, dan bimbingan dari berbagai pihak. Semoga segala bantuan, bimbingan, dan pengajaran yang telah diberikan kepada penulis mendapatkan imbalan dari Allah SWT. Penulis mohon maaf apabila selama proses penyusunan tugas akhir ini terdapat kekhilafan dan kesalahan. Penulis menyadari sepenuhnya keterbatasan kemampuan dalam penulisan tugas akhir ini, oleh karena itu penulis mengharap adanya kritik dan saran yang membangun demi kesempurnaan penyusunan dan penulisan tugas akhir ini. Akhir kata, semoga tugas akhir ini dapat bermanfaat bagi semua yang membaca dan membutuhkan, Aamiin aamiin ya robbal'alamiin.

Wassalamu'alaikum, Wr. Wb

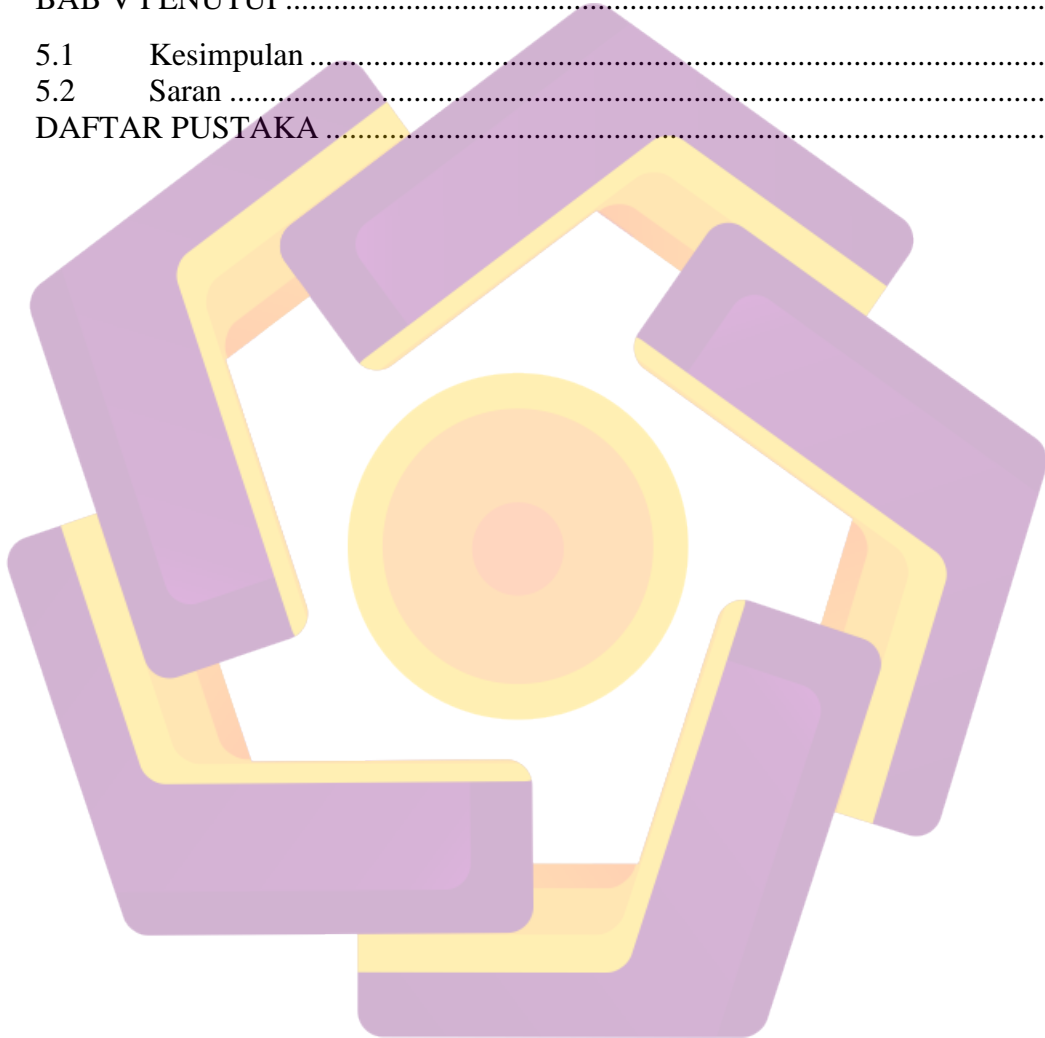
Yogyakarta, 20 Februari 2023

Penulis

DAFTAR ISI

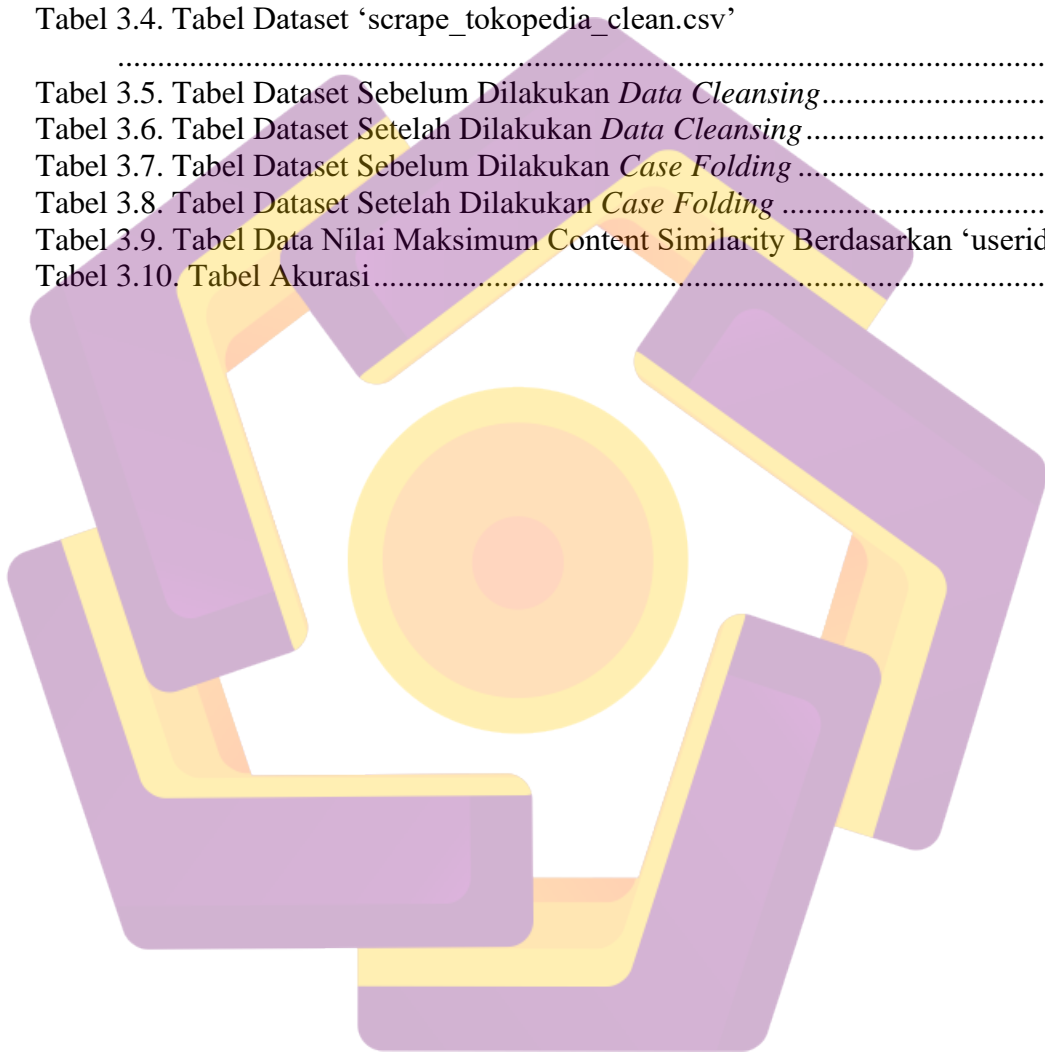
HALAMAN JUDUL.....	i
HALAMAN PERSETUJUAN.....	iii
HALAMAN PENGESAHAN.....	iv
HALAMAN PERNYATAAN KEASLIAN SKRIPSI	v
HALAMAN PERSEMBAHAN	vi
KATA PENGANTAR	vii
DAFTAR ISI.....	viii
DAFTAR TABEL.....	x
DAFTAR GAMBAR	xi
INTISARI.....	xiii
<i>ABSTRACT</i>	xiv
BAB I PENDAHULUAN	1
1.1 Latar Belakang Masalah	1
1.2 Rumusan Masalah.....	2
1.3 Batasan Masalah	3
1.4 Tujuan Penelitian	4
1.5 Manfaat Penelitian	4
1.6 Sistematika Penulisan	5
BAB II LANDASAN TEORI	6
2.1 Tinjauan Pustaka.....	6
2.2 Ulasan	12
2.3 <i>Machine Learning</i>	12
2.4 <i>Natural Language Processing</i>	13
2.5 <i>Logistic Regression</i>	14
2.6 Python	16
2.7 TF-IDF	17
2.8 Pickle.....	19
2.9 Selenium-Wire	21
2.10 <i>Evaluation Measurement</i>	21
BAB III METODOLOGI PENELITIAN	23
3.1 Analisis Masalah.....	23
3.2 Solusi Yang Dipilih.....	24
3.3 Alur Penelitian	26
3.4 Alat dan Bahan.....	27
3.5 Metodologi Penelitian.....	27

BAB IV PEMBAHASAN.....	45
4.1 Pengumpulan Dataset.....	45
4.2 <i>Preprocessing Data</i>	54
4.3 <i>Data Training</i>	67
4.4 <i>Data Testing</i>	71
4.5 <i>Import Model</i>	75
4.6 Implementasi Model	76
BAB V PENUTUP.....	80
5.1 Kesimpulan	80
5.2 Saran	80
DAFTAR PUSTAKA	82



DAFTAR TABEL

Tabel 2.1. Tabel Perbandingan Studi Pustaka.....	9
Tabel 3.1. Tabel Dataset ‘train_shopee.csv’	29
Tabel 3.2. Tabel Klasifikasi Dataset ‘train_shopee.csv’	30
Tabel 3.3. Tabel Dataset ‘scrape_shopee_clean.csv’	30
Tabel 3.4. Tabel Dataset ‘scrape_tokopedia_clean.csv’	31
Tabel 3.5. Tabel Dataset Sebelum Dilakukan <i>Data Cleansing</i>	33
Tabel 3.6. Tabel Dataset Setelah Dilakukan <i>Data Cleansing</i>	33
Tabel 3.7. Tabel Dataset Sebelum Dilakukan <i>Case Folding</i>	34
Tabel 3.8. Tabel Dataset Setelah Dilakukan <i>Case Folding</i>	35
Tabel 3.9. Tabel Data Nilai Maksimum Content Similarity Berdasarkan ‘userid’	36
Tabel 3.10. Tabel Akurasi.....	37



DAFTAR GAMBAR

Gambar 2.1. Rumus Logistic Regression.....	15
Gambar 2.2. Rumus TF-IDF	18
Gambar 2.3. Rumus TF.....	19
Gambar 2.4. Rumus IDF	19
Gambar 2.5. Pemetaan <i>confusion matrix</i>	22
Gambar 3.1. Alur Penelitian.....	26
Gambar 3.2. <i>Flowchart Program Scraping data</i>	41
Gambar 3.3. <i>Flowchart website</i>	42
Gambar 3.4. <i>Activity Diagram Website</i>	43
Gambar 3.5. Desain Antarmuka <i>Website</i>	44
Gambar 4.1. <i>Library</i> yang digunakan pada program <i>web scrape</i> Shopee	46
Gambar 4.2. Fungsi ‘getUrl’ dan Fungsi ‘getAccess’	47
Gambar 4.3. Fungsi utama yang menjalankan program	47
Gambar 4.4. Hasil Scrape Program.....	48
Gambar 4.5. Program ‘bersih_data.py’	49
Gambar 4.6. Hasil Program ‘bersih_data.py’	50
Gambar 4.7. Fungsi ‘getParam()’	51
Gambar 4.8. Fungsi ‘getPayload()’ dan Fungsi ‘scrapeData()’	52
Gambar 4.9. Fungsi ‘main’	53
Gambar 4.10. Program ‘bersih_data.py’ untuk Tokopedia.....	54
Gambar 4.11. Hasil Program ‘bersih_data.py’ Tokopedia	54
Gambar 4.12. Import Data ‘scrape_shopee_clean.csv’	55
Gambar 4.13. Data Sebelum Dilakukan <i>Labeling</i>	56
Gambar 4.14. Data Setelah Dilakukan <i>Labeling</i>	56
Gambar 4.15. Halaman Awal Jupyter Notebook	58
Gambar 4.16. Import Library yang Dibutuhkan	58
Gambar 4.17. Import ‘dataframe’ dengan Data dari ‘train_shopee.csv’ dan ‘train_tokopedia.csv’ Serta Menampilkan Lima Data Pertama.....	60
Gambar 4.18. Menampilkan Deskripsi Data untuk Setiap Kolom Dataframe.....	60
Gambar 4.19. Menampilkan Total Data yang Kosong (bernilai null) dan menghapusnya	61
Gambar 4.20. Menghapus Kolom ‘orderid’ karena Memiliki Data Kosong serta Menampilkan Jumlah Data untuk Setiap Nilai dari Kolom ‘fakeornot’.....	61
Gambar 4.21. Proses <i>Data Cleansing, Case Folding, Tokenization</i>	62
Gambar 4.22. Proses Mengubah <i>Unix Timestamp</i> Menjadi Tanggal dan Waktu serta mengelompokkannya sesuai dengan id pengguna.....	63
Gambar 4.23. Mencari <i>Cosine Similarity</i> dan Membentuk Diagonal Matriks	64
Gambar 4.24. Output dari ‘df3’	65
Gambar 4.25. Data ‘df3’ Dilakukan Merge ke Dataframe Utama.....	66
Gambar 4.26. Menghapus Kolom pada Dataframe ‘train_shopee.csv’ yang tidak Digunakan.....	66
Gambar 4.27. Menghapus Kolom pada Dataframe ‘scrape_shopee_clean.csv’ yang tidak Digunakan	66
Gambar 4.28. Data <i>Training</i>	68

Gambar 4.29. <i>Confusion Matrix</i>	69
Gambar 4.30. Visualisasi <i>Confusion Matrix</i>	69
Gambar 4.31. Hasil Akurasi, Presisi, <i>Recall</i> , <i>F1-score</i> Pada Data ‘train_shopee.csv’	70
Gambar 4.32. Hasil Perbandingan Ulasan <i>Fake</i> dengan Ulasan Original Pada Data ‘train_shopee.csv’	70
Gambar 4.33. Data Testing ‘scrape_shopee_clean.csv’	71
Gambar 4.34. Output Jumlah <i>Fake</i> dan Original	72
Gambar 4.35. Menampilkan Visualisasi Data dalam bentuk Persentase	72
Gambar 4.36. Output Visualisasi Ulasan Palsu dan Asli dalam Bentuk Persentase	73
Gambar 4.37. Menampilkan Visualisasi Sebaran Ulasan Palsu dalam Bentuk Histogram	73
Gambar 4.38. Visualisasi Sebaran Panjang Ulasan palsu	74
Gambar 4.39. Menampilkan Visualisasi Sebaran Ulasan Asli dalam Bentuk Histogram	74
Gambar 4.40. Visualisasi Sebaran Panjang Ulasan Asli.....	75
Gambar 4.41. Assign Variable model dengan fungsi <i>LogisticRegression</i>	75
Gambar 4.42. Import Model dengan Nama ‘model.pickle’	76
Gambar 4.43. Mengambil Dataset dan Menghapus kolom yang tidak digunakan pada Dataset Shopee	76
Gambar 4.44. Halaman Utama Website	77
Gambar 4.45. Proses Pendeteksian	78
Gambar 4.46. Hasil Deteksi	79

INTISARI

Di era yang serba instan ini, proses belanja daring sudah menjadi hal yang biasa. Mulai dari kebutuhan pokok hingga alat-alat yang dapat dipakai sehari-hari, barang-barang ini dapat didapatkan secara daring dengan cara yang sangat mudah. Namun dewasa ini, banyak pembeli yang membeli barang dengan hati-hati, khawatir akan penipuan, pembeli biasanya melihat deskripsi terlebih dahulu. Hal tersebut biasa dilakukan agar menghindari penipuan kualitas, yaitu barang yang tampil di *e-commerce* tidak sesuai dengan barang yang datang, dan hal itu sering terjadi. Ada juga hal yang biasa pembeli lakukan adalah melihat ulasan yang disediakan di toko tersebut. Ulasan yang dilihat dinilai dari positif atau negatifnya, sehingga tidak sedikit penjual yang memalsukan ulasan agar mendapat *insight* negatif dari pembeli. Maka dari itu saya akan melakukan penelitian dengan judul “Deteksi Ulasan Palsu pada Tokopedia dan Shopee Menggunakan *Natural Language Processing*” dengan bahasa pemrograman yang digunakan python berbasis analisis.

Solusi yang akan digunakan pada penelitian ini berdasarkan analisa masalah yang sudah diuraikan sebelumnya yaitu membuat system berbasis website yang dapat mendeteksi apakah ulasan tersebut dapat dikatakan palsu atau tidak. Sebagai gambaran ulasan tersebut dapat dikatakan palsu yang bersumber dari blog Blibli.com setidaknya dia memiliki 4 hal yaitu ulasan yang tidak hanya rating-nya saja melainkan ada komentarnya, ulasan yang menggunakan bahasa yang berlebihan besar kemungkinan ulasan tersebut palsu, ulasan kemungkinan besar tidak berada pada toko yang sama dengan jam dan tanggal yang sama, lalu yang terakhir adalah dengan melihat akun pengulas apakah pengulas melakukan ulasan yang banyak dalam waktu banyak atau tidak. Untuk parameter terakhir pada Tokopedia dan Shopee tampaknya tidak memungkinkan untuk dilakukan sehingga parameter yang diambil hanya 3 bagian pertama, karena untuk pengecekan akun pengulas akan membutuhkan data pengguna yang cukup banyak, itupun sangat sulit karena data tersebut merupakan data sensitif.

Berdasarkan proses tersebut hingga implementasi, dapat disimpulkan bahwa: hasil penelitian menghasilkan website yang dapat mendeteksi ulasan palsu yang diinputkan berdasarkan dataset yang telah dilakukan input serta tingkat akurasi model ini karena memiliki dataset yang identik pada kedua platform melalui proses training memiliki hasil score yang baik, baik itu dari confusion matrix, akurasi, presisi, recall, serta f1-score.

Kata kunci: *machine learning, natural language processing, ulasan palsu, python*

ABSTRACT

In this instant era, online shopping has become commonplace. From basic necessities to tools that can be used every day, these items can be obtained online in a very easy way. But today, many buyers buy goods carefully, worried about fraud, buyers usually look at the description first. This is usually done in order to avoid quality fraud, namely the goods that appear in e-commerce do not match the goods that come, and this often happens. There is also something that buyers usually do is look at the reviews provided in the store. The reviews that are seen are judged on the positive or negative side, so not a few sellers fake reviews to get negative insights from buyers. Therefore, I will conduct research with the title "Detecting Fake Reviews on Tokopedia and Shopee Using Natural Language Processing" with the programming language used python-based analysis.

The solution that will be used in this study is based on the analysis of the problems previously described, namely creating a website-based system that can detect whether the review can be said to be fake or not. As an illustration, this review can be said to be fake, sourced from the Blibli.com blog, at least it has 4 things, namely reviews that are not only rated but have comments, reviews that use excessive language are likely to be fake, reviews are most likely not on the same store with the same hours and the same date, then the last thing is to look at the reviewer's account whether the reviewer did a lot of reviews in a lot of time or not. For the last parameter on Tokopedia and Shopee it seems impossible to do so only the first 3 parameters are taken, because checking the reviewer's account will require quite a lot of user data, and even then it is very difficult because the data is sensitive data.

From this process to implementation, it can be concluded that: the research results produce a website that can detect fake reviews that are inputted based on the dataset that has been inputted and the level of accuracy of this model because it has identical datasets on both platforms through the training process has a good score, both that's from the confusion matrix, accuracy, precision, recall, and f1-score.

Keyword: *machine learning, natural language processing, fake reviews, python*