Penggunaan Regular Expression dalam Penambangan Data Email untuk Deteksi Phishing

Evelyn Yosiana - 1352083
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jalan Ganesha 10 Bandung
E-mail (gmail): 13522083@std.stei.itb.ac.id

Abstrak. *Phishing* merupakan ancaman keamanan siber (*cyber* security) di mana penyerang mencoba memperoleh informasi sensitif seperti username, password, dan informasi kartu kredit dengan menyamar sebagai lembaga terpercaya yang salah satunya melalui email. Makalah ini membahas tentang penggunaan *regular expression* (regex) dalam deteksi email *phishing.* Dalam makalah ini terdapat penjelasan mengenai konsep dasar *regular expression*, implementasi dalam bahasa pemrograman Python, dan pengujian akurasi menggunakan dataset dari Kaggle. Hasil analisis menunjukkan bahwa penggunaan Regex dapat menjadi alat yang efektif untuk mendeteksi email phishing, namun masih memerlukan optimalisasi lebih lanjut untuk meningkatkan akurasi deteksi.

Kata kunci: *regular expression*, email, *phishing*.

I. PENDAHULUAN

Phishing merupakan salah satu ancaman keamanan siber (*cyber security*) yang cukup umum ditemui, di mana penyerang mencoba untuk memperoleh informasi seperti username, password, dan informasi kartu kredit dengan menyamar sebagai entitas yang terpercaya dalam komunikasi elektronik. Salah satu metode paling umum yang digunakan dalam serangan *phishing* adalah melalui email. Seiring dengan meningkatnya jumlah pengguna internet dan email, ancaman phishing semakin berkembang dan menjadi lebih canggih. Oleh karena itu, deteksi dan mitigasi serangan *phishing* menjadi salah satu hal penting dalam keamanan siber. Salah satu teknik yang efektif untuk mendeteksi email *phishing* yaitu dengan menggunakan *regular expression* (regex).

Makalah ini bertujuan untuk menjelaskan bagaimana Regex dapat digunakan sebagai alat untuk mendeteksi email *phishing*. Dalam makalah ini, penulis akan membahas konsep dasar regex, cara kerja regex dalam mendeteksi pola tertentu yang umum ditemukan dalam email phishing, serta implementasi praktis menggunakan bahasa pemrograman python. Untuk mengetes tingkat keakuratan dari implementasi yang telah dibuat, Penulis juga akan membandingkan hasilnya dengan suatu dataset yang diperoleh dari Kaggle.

II. TEORI DASAR

A. *Regular Expression*

Regular Expression (Regex) adalah suatu pola yang digunakan untuk mencocokkan karakter dalam sebuah string. Regex terdiri dari berbagai komponen dan notasi yang memungkinkan pengguna untuk membangun pola pencarian yang kompleks. Beberapa komponen dasar Regex antara lain:

- **Karakter biasa**: karakter biasa seperti huruf dan angka yang dicari secara langsung dalam teks.
- **Metakarakter**: simbol khusus yang memiliki makna khusus dalam Regex. Beberapa metakarakter penting antara lain sebagai berikut.

  . : mewakili karakter apa saja kecuali newline.

  \d : mewakili digit (0-9).

  \w : mewakili karakter kata (alphanumeric dan underscore).

  \s : mewakili spasi (whitespace).

Pola dasar dalam regex antara lain sebagai berikut.

- **Kelas Karakter**: digunakan untuk mencocokkan salah satu dari beberapa karakter.

  [abc] : Mencocokkan 'a', 'b', atau 'c'.

  [^abc] : Mencocokkan karakter selain 'a', 'b', atau 'c'.

  [a-z] : Mencocokkan semua huruf kecil dari 'a' hingga 'z'.
- **Kuantisasi**: digunakan untuk menentukan jumlah kemunculan dari elemen sebelumnya.

  a* : Mencocokkan 'a' 0 kali atau lebih.

  a+ : Mencocokkan 'a' 1 kali atau lebih.

  a? : Mencocokkan 'a' 0 atau 1 kali.

  a{n} : Mencocokkan tepat sejumlah n 'a'.

  a{n, m} : Mencocokkan antara n hingga m 'a'.

B. *Phishing*

Menurut S. Jeeva, *phishing* merupakan tindakan kriminal online yang terjadi ketika sebuah halaman web berbahaya menyamar sebagai halaman web yang sah untuk mendapatkan informasi sensitif dari pengguna. Dilansir dari *oxford dictionary*, kegiatan *phishing* berarti "the activity of tricking people by getting them to give their identity, bank account numbers, etc. over the internet or by email, and then using these to steal money from them" atau dalam bahasa Indonesia "kegiatan menipu orang dengan membuat mereka memberikan identitas, nomor rekening bank, dll. melalui internet atau email, dan kemudian menggunakannya untuk mencuri uang dari mereka". *Phishing* sendiri memiliki berbagai dampak yang cukup merugikan, antara lain sebagai berikut.

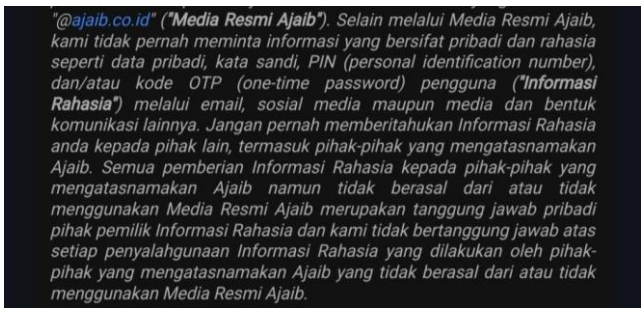
- Kerugian Finansial: pencurian informasi keuangan dapat menyebabkan kerugian uang dalam jumlah besar.
- Kehilangan data: akses ke data sensitif yang dapat merugikan privasi seseorang atau sekelompok orang.
- Kerugian reputasi: terungkapnya data seseorang atau sekelompok orang dapat merusak reputasi orang atau sekelompok orang tersebut serta mengurangi kepercayaan pelanggan.

C. *Phishing* pada Email

Menurut sebuah studi di Frontiers in Psychology, email phishing sering kali manipulasi emosional dan mendorong seseorang untuk bertindak dengan cepat untuk menipu penerima. Berikut beberapa contoh kata yang berpotensi menjadi indikasi email *phishing* dengan manipulasi emosi.

- "urgent action required"
- "immediate attention needed"
- "your account will be suspended"

Perlu diingat bahwa lembaga atau organisasi resmi jarang, bahkan tidak pernah meminta informasi sensitif melalui email.



Berikut beberapa contoh kata yang berpotensi menjadi indikasi email *phishing* dengan meminta data yang sifatnya privasi.

- "confirm your identity"
- "verify your account"
- "update your payment information"

Lembaga dan organisasi resmi juga jarang menyertakan ajakan untuk mengeklik tautan tertentu.

Berikut beberapa contoh kata yang berpotensi menjadi indikasi email *phishing* dengan tautan.

- “click here”
- “download now”
- “open this attachment”

Selain itu, email *phishing* terkadang mengandung kesalahan kata ataupun tata bahasa. Hal ini dapat mengindikasikan bahwa email yang dikirimkan tidak ditulis oleh organisasi atau lembaga yang profesional.

III. ANALISIS DAN PEMBAHASAN

Untuk mendeteksi email *phishing*, pertama-tama perlu diketahui kata-kata apa saja yang dapat mengindikasikan bahwa suatu email tergolong *phishing*.

A. Mendefinisikan pola *regular expression*

Pertama-tama didefinisikan pola *regular expression* yang akan digunakan. Pendefinisian ini berdasarkan pada hasil *reseacrh* yang dilakukan oleh penulis. Berikut pola-pola *regular expression* pertama yang akan digunakan.

- .urgent.
- .immediate(ly)?.
- .important.
- .now.
- .required?s?.
- .needed.
- .alert.
- .verify?(ication)?.
- .confirm(ation?).
- .update.
- .secure?(ity)?.
- .account.
- .password.
- .identity.
- .susp(icious)?(ended)?.
- .fraudulent.
- .compromised.
- .detected.
- .login.
- .click.
- .download.
- .open.
- .setting.
- .attachment.
- .offer.
- .unauthorized.
- .(re)?activate.

B. Pengecekan dengan Dataset

Pengecekan keakuratan pola yang digunakan dapat dilakukan dengan beberapa cara. Pada makalah ini, penulis menggunakan sebuah dataset dari Kaggle. Dataset ini terdiri dari 18650 data yang terdiri dari “Email Text” dan “Email Type” dimana “Email Type” sendiri terdiri atas “Phishing Email” dan “Safe Email”.

Pengecekan dimulai dengan memuat dataset kemudian menduplikat isinya ke dalam sebuah *list of dictionary* dimana *key* dari *dictionary* tersebut berisi “Email Text” dan *value*-nya berisi “Email Type”. Untuk setiap *dictionary* akan dimasukkan ke sebuah fungsi untuk mengecek apakah terdapat pola *regular expression* yang telah didefinisikan sebelumnya di dalam teks email tersebut. Jika ada, maka email tersebut diprediksi sebagai *phishing email*.

Setelah tiap email selesai diprediksi, untuk tiap pola yang telah didefinisikan, akan dihitung berapa kali pola tersebut menyebabkan email yang sebenarnya aman terdeteksi *phishing*. Berikut hasil pada terminal.

```
\burgent\b: 63
\bimmediate(ly)?\b: 416
\bimportant\b: 640
\bnow\b: 1887
\brequired?s?\b: 880
\bneeded\b: 366
\balert\b: 48
\bverify?(ication)?\b: 145
\bconfirm(ation?)?\b: 278
\bupdate\b: 584
\bsecure?(ity)?\b: 408
\baccount\b: 414
\bpassword\b: 122
\bidentity\b: 108
\bsusp(icious)?(ended)?\b: 41
\bfraudulent\b: 5
\bcompromised\b: 22
\bdetected\b: 215
\blogin\b: 54
\bclick\b: 718
\bdownload\b: 242
\bopen\b: 658
\bsetting\b: 198
\battachment\b: 80
\boffer\b: 448
\bunauthorized\b: 47
\b(re)?activate\b: 15
```

```
Sebenarnya safe email namun terprediksi phishing:
5147 / 11322
Sebenarnya phishing email namun terprediksi safe:
2785 / 7328

Akurasi: 0.5746916890080429
```

C. *Optimalisasi*

Pada makalah ini, penulis melakukan optimalisasi pola dengan mencari frasa yang terdapat pada *phishing email* namun tidak terdapat pada *safe email* dengan jumlah tertentu.

Untuk setiap kata yang banyak digunakan di *phishing email* maupun *safe email*, akan dicari kata sebelumnya dan kata sesudahnya dalam teks email untuk digabungkan menjadi dua buah frasa menggunakan sebuah fungsi. Hal tersebut dilakukan pada email *safe* di dataset dan email *phishing* di dataset. Setelah itu, untuk tiap daftar frasa di email *phishing* akan dicocokan dengan daftar frasa di email *safe*. Fungsi ini akan mengembalikan tiap frasa yang terdapat pada email *phishing* namun tidak terdapat pada email *safe* beserta jumlahnya (secara terurut mengecil).

Untuk kata-kata yang dicurigai namun membuat email yang bukan *phishing* terdeteksi menjadi email *phishing*, dapat dimasukkan ke kode tersebut, kemudian diambil hasilnya dengan jumlah yan cukup besar untuk kemudian dimasukkan ke dalam pola. Pengambilan hanya dilakukan pada frasa yang paling sering muncul untuk menghindari *overfit* terhadap dataset. Berikut pola yang telah dioptimalisasi oleh penulis.

- .urgent.
- .prescription.required.
- .required.(!|input).
- .immediate.download.
- .immediately.sell.
- .action.may.
- .(before|viagra).action.
- .alert.
- .update.your.information.
- .security.check.
- .verify.your.account.
- .confirm.your.password.
- .suspicious.
- .fraudulent.
- .compromised.
- .login.
- .click.here.
- .[A-Za-z0-9._%+-]{100,}.
- .attachment.
- .unauthorized.
- .reactivate.
- .[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+.(com|co|org).
- .http.
- .provide.your.credit.card.
- .debit.
- .remove.
- .for.free.
- .best.deal.
- .prize.
- .unbelievable.
- .special.promotion.
- .porn.
- .sex.
- .hot.girls.
- .horny.
- .xxx.
- .sign.up.
- .grow.up.to.
- .maillist.verify.
- .verify.now.
- .(we|combination|intro|final).offer.
- .offer.(latest|because|localized|includes|creative|refer|#|fair|going|manager|unsubscribe|only|--).
- .(material|right|running|download|order|filings|logo).now.
- .now.(qualify|deposited|for|furthered).
- .(duty|a|100%|in|obtaining|placing|includes?|hitch).free.
- .free.(zone|trading|zonedubai|bonus|grants?|quote|instant|personal|lifetime|application|government|no|quote|shipping|go).

Berikut hasil optimalisasi yang telah dilakukan oleh penulis.

```
\burgent\b : 63
\bprescription\s*required\b : 0
\brequired\s*(!|input)\b : 0
\bimmediate downioad\b : 0
\bimmediately\s*sell\b : 0
\baction\s*may\b : 0
\b(before|viagra)\s*action\b : 0
\balert\b : 48
\bverify\b : 0
\bupdate\s*your\s*information\b : 1
\bsecurity\s*check\b : 3
\bverify\s*your\s*account\b : 0
\bconfirm\s*your\s*password\b : 0
\bsuspicious\b : 23
\bfraudulent\b : 5
\bcompromised\b : 22
\blogin\b : 54
\bclick\s*here\b : 107
\b[A-Za-z0-9._%+-]{100,}\b : 3
\battachment\b : 80
\bunauthorized\b : 47
\breactivate\b : 2
\b[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+\.^(com|co|org)\b : 0
```

```
\bhttp\b : 0
\bprovide\s*your\s*credit\s*card\b : 0
\bdebit\b : 16
\bremove\b : 189
\bfor\s*free\b : 165
\bbest\s*deal\b : 4
\bprize\b : 41
\bunbelievable\b : 6
\bspecial\s*promotion\b : 1
\bporn\b : 23
\bsex\b : 79
\bhot\s*girls\b : 1
\bhorny\b : 1
\bxxx\b : 14
\bsign\s*up\b : 97
\bgrow\s*up\s*to\b : 1
\b(we|combination|intro|final)\s*offer\b : 8
\boffer\s*(latest|because|localized|includes|creative|refer|#|fair|going|manag
er|unsubscribe|only|--)\b : 0
\b(material|right|running|download|order|filings|logo)\s*now\b : 174
\bnow\s*(qualify|deposited|for|furthered)\b : 26
\bmaillist\s*verify\b : 0
\bverify\s*now\b : 0
\b(duty|a|100%|in|obtaining|placing|includes?|hitch)\s*free\b : 90
\bfree\s*(zone|trading|zonedubai|bonus|grants?|quote|instant|personal|lifetime
|application|goverment|no|quote|shipping|go)\b : 17
```

```
Sebenarnya safe email namun terprediksi phishing:
1120 / 11322
Sebenarnya phishing email namun terprediksi safe:
3924 / 7328

Akurasi: 0.729544235924933
```

D. Kasus Uji

Untuk menerima input dari pengguna, penulis juga membuat program main dengan input berupa teks email (string). Berikut comtoh hasil eksekusinya.

Kasus uji dari email asli:

Selamat malam peserta mata kuliah IF2211 Strategi Algoritma.

Melalui milis ini kami ingin mengumumkan bahwa jadwal demo tugas besar 3 sudah dapat diisi.

Penjadwalan demo dapat dilakukan melalui tautan bit.ly/kelompoktubes3stima24. Demo wajib sudah **dijadwalkan sebelum Minggu, 9 Juni 23.59 WIB**. Bagi kelompok yang tidak mengisi hingga batas yang ditentukan maka komponen demo kelompok tersebut akan menjadi nol.

Atas perhatiannya, kami ucapkan terimakasih.

Salam,

```
Masukkan teks email: Penjadwalan demo dapat dilakukan melalui
tautan bit.ly/kelompoktubes3stima24. Demo wajib sudah dijadwal
kan sebelum Minggu, 9 Juni 23.59 WIB. Bagi kelompok yang tidak
 mengisi hingga batas yang ditentukan maka komponen demo kelom
pok tersebut akan menjadi nol.
Email aman.
```

Berikut adalah Petunjuk Pelaksanaan UAS IF2220 Probabilitas dan Statistika. Harap diperhatikan dan diikuti dengan baik untuk kelancaran pelaksanaan ujian.

1. UAS dilaksanakan secara luring pada hari Senin, 3 Juni 2024 mulai pukul 12.30 WIB.
2. Materi ujian: Distribusi Sampel, Estimasi, Uji Hipotesis, Regresi linear dan Korelasi. Materi sebelumnya tidak diujikan khusus, tapi tetap tercakup dalam ujian.
3. Sifat ujian: Individual, Closed book, Open note (1 lembar A4 cheat sheet, tulisan tangan sendiri), boleh menggunakan Kalkulator Non-gadget.
4. Tabel statistika akan disediakan di lembar soal.
5. Memakai pakaian sopan dan sepatu, dilarang menggunakan sandal.

Terima kasih,

Tim Dosen IF2220

```
Masukkan teks email: Berikut adalah Petunjuk Pelaksanaan UAS I
F2220 Probabilitas dan Statistika. Harap diperhatikan dan diik
uti dengan baik untuk kelancaran pelaksanaan ujian.1. UAS dila
ksanakan secara luring pada hari Senin, 3 Juni 2024 mulai puku
l 12.30 WIB.2. Materi ujian: Distribusi Sampel, Estimasi, Uji
Hipotesis, Regresi linear dan Korelasi. Materi sebelumnya tida
k diujikan khusus, tapi tetap tercakup dalam ujian.3. Sifat uj
ian: Individual, Closed book, Open note (1 lembar A4 cheat she
et, tulisan tangan sendiri), boleh menggunakan Kalkulator Non-
gadget.4. Tabel statistika akan disediakan di lembar soal.5. M
emakai pakaian sopan dan sepatu, dilarang menggunakan sandal.
Email aman.
```

Kasus uji dari database:

```
is the number one adult content provider . thank you for your time . scarlett",Phishing Email
```

```
Masukkan teks email: premium adult content looking for high qu
ality adult content at the right price ? then check out xxx by
 scarlett - 15 image cd 's and 3 video clip cd 's * hardcore *
 softcore * asian hardcore * transsexual hardcore * english ro
ses ( 18-21 yrs ) * extreme euro hardcore * gay * amateur chec
k out our quality , the amazing prices and with new content co
ming out every month you ' ll easily see why xxx by scarlett i
s the number one adult content provider . thank you for your t
ime . scarlett
Hati-hati, email terdeteksi phishing!
```

```
additional feedback to : service dept 9420 reseda blvd # 133 northridge , ca 91324",Phishing Email
```

```
Masukkan teks email: "free portable dvd player offer valid onl
y to residents of the united states who are at least 18 years
old . you need to complete our offer eligibility requirements
to qualify for your free gift . trademarks , service marks , l
ogos , and / or domain names are the property of their respect
ive owners , who have no association with or make any endorsem
ent of the products or services provided by lookdog . com . to
 unsubscribe from future lookdog . com promotions and offers p
lease submit an online request using the link below . you may
also send a written request to : 3250 w . big beaver road , su
ite 144 , troy , mi , 48084 . click here to unsubscribe . plea
se refer all questions , opinions or additional feedback to :
service dept 9420 reseda blvd # 133 northridge , ca 91324"
Hati-hati, email terdeteksi phishing!
```

Untuk *source code* lengkapnya dapat dilihat melalui tautan berikut:
https://github.com/evelynnn04/regex_for_detecting_phishing_email.git

IV. KESIMPULAN

Penelitian ini menunjukkan bahwa *regular expression* dapat digunakan sebagai alat yang efektif untuk mendeteksi email *phishing*. Implementasi pola-pola *regular expression* yang digunakan dalam penelitian ini berhasil mengidentifikasi email *phishing* berdasarkan dataset yang digunakan. Namun, hasil penelitian juga menunjukkan adanya beberapa kesalahan deteksi, baik dalam mengidentifikasi email yang aman sebagai *phishing* maupun sebaliknya. Optimalisasi lebih lanjut pada pola-pola *regular expression* dapat meningkatkan akurasi deteksi.

V. LINK VIDEO

https://youtu.be/L4Tv8IIO5XQ?si=f-NapI7aK2wBk7dP

VI. UCAPAN TERIMA KASIH

Pada kesempatan ini, penulis mengucapkan terima kasih kepada Tuhan Yang Maha Esa karena berkat rahmatnya penulis dapat menyelesaikan makalah ini dengan baik dan lancar, sesuai dengan batas waktu yang telah ditetapkan. Selain itu, penulis juga ingin mengucapkan terima kasih kepada pihak-pihak terkait yang telah membantu penulis dalam menyelesaikan makalah ini:

- bapak Dr. Ir. Rinaldi Munir, M.T. selaku dosen pengampu mata kuliah strategi algoritma yang telah membimbing penulis selama satu semester penuh dalam mata kuliah ini;
- orangtua penulis yang selalu mendukung, memberi semangat, dan mendoakan penulis;
- teman-teman penulis selaku *support system* penulis dalam menyelesaikan makalah ini; serta
- asisten mata kuliah strategi algoritma yang telah membimbing tugas-tugas (baik tugas besar maupun tugas kecil) mata kuliah strategi algoritma.

VII. REFERENSI

[1] phishing noun - Definition, pictures, pronunciation and usage notes | Oxford Advanced Learner's Dictionary at OxfordLearnersDictionaries.com. (n.d.). https://www.oxfordlearnersdictionaries.com/definition/english/phishing?q=phishing

[2] Jeeva, S. C., & Rajsingh, E. B. (2016). Intelligent phishing url detection using association rule mining. *Human-centric Computing and Information Sciences*, *6*(1). https://doi.org/10.1186/s13673-016-0064-3

[3] Munir, Rinaldi. (2024). String Matching dengan Regex: Bandung. https://informatika.stei.itb.ac.id/~rinaldi.munir/Stmik/2022-2023/String-Matching-dengan-Regex-2019.pdf

[4] Munir, Rinaldi. (2024). Modul Praktikum Kuliah, Pengantar Regular Expression: Bandung. https://informatika.stei.itb.ac.id/~rinaldi.munir/Stmik/2019-2020/Modul-Praktikum-NLP-Regex.pdf

[5] Elnahas. (2023, September 21). *Phishing Email Detection using SVM & RFC.* https://www.kaggle.com/code/elnahas/phishing-email-detection-using-svm-rfc/input

[6] *Frontiers in Psychology*. (2021). Phishing Attacks: A Recent Comprehensive Study and a New Anatomy. https://www.frontiersin.org/articles/10.3389/fcomp.2021.563060/full

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 12 Juni 2024

Evelyn Yosiana

13522083