Pencegahan Pelecehan Seksual pada Media Sosial

Menggunakan Metode Ekstraksi Kata Kunci pada Teks dan *Regular Expression*

Muhammad Rahadian Alamsyah Putra Winarno 13518011
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jalan Ganesha 10 Bandung
E-mail : rahadian.apw@gmail.com

***Abstraksi*—Media sosial adalah media daring yang dapat digunakan manusia untuk melakukan kegiatan sosial. Tidak dapat dipungkiri bahwa saat ini media sosial sangat penting pada berbagai bidang. Media sosial tidak hanya digunakan untuk sekadar melepas rindu namun juga digunakan untuk pendidikan, sosialisasi, bisnis, dan lainnya. Seiring berkembangnya teknologi media sosial juga memiliki fitur yang berkembang juga. Namun sangat disayangkan bentuk kejahatan sosial seperti pelecehan seksual pun berkembang pada media ini. Pelecehan seksual merupakan segala kegiatan yang memiliki arah menuju seksual dan mengganggu kenyamanan. Hal ini dapat dicegah dengan mendeteksi pesan yang dikirim menggunakan algoritma pencocokan string yaitu *regular expression* agar pesan yang mengandung ujaran pelecehan seksual dapat dihentikan.**

Kata Kunci—kata kunci; kecerdasan buatan; media sosial; pelecehan; pembelajaran mesin; regular expression;

I. Pendahuluan

Baru-baru ini marak terjadi pelecehan seksual di media sosial. Sangat disayangkan masih banyak orang yang memanfaatkan kemajuan teknologi tanpa berempati terhadap pengguna lainnya. Bahkan kasus ini terjadi didalam lingkungan kampus yang isinya adalah orang yang seharusnya berpendidikan. Kejahatan seperti ini tidak mengenal lingkungan, taraf pendidikan, bahkan agama selama masih ada orang yang hanya memikirkan diri sendiri tanpa memikirkan keadaan orang lain.

Korban pelecehan seksual pada umumnya tidak dapat melakukan perlawanan dan hanya menurut karena korban sedang bingung dengan apa yang terjadi. Kemudian terjadilah kasus pelecehan yang mengakibatkan dampak terhadap korban. Pola yang dilakukan oleh pelaku dapat dikenali oleh karena itu makalah ini dibuat.

Media sosial pada umumnya melakukan pelayanan komunikasinya menggunakan pesan teks. Pesan ini terdiri dari kumpulan kata atau karakter yang dapat diolah menjadi informasi pesan. Pengolahan kata dapat dilakukan dengan menggunakan algoritma pencocokan string sehingga didapatkan maksud dari kumpulan kata yang terdapat pada pesan tersebut.

II. Dasar Teori

A. Media Sosial

Media sosial adalah laman atau aplikasi yang memungkinkan pengguna untuk melakukan kegiatan sosial didalamnya. Kegiatan sosial bisa berupa berbincang-bincang melalui video, suara, atau text, membagikan momen menarik, berbagi cerita, berbagi pengalaman, dll. Media sosial tercipta karena adanya teknologi internet. Dengan media sosial orang bisa berkomunikasi tanpa mengenal jarak dan waktu.

Gambar 1. Media Sosial sumber: play.google.com

B. Pelecehan Seksual

Pelecehan seksual adalah perilaku atau perhatian yang bersifat seksual dan tidak diinginkan korban atau dapat membuat korban terganggu. Pelecehan pada media sosial dilakukan secara verbal baik secara implisit maupun eksplisit. Pola yang biasanya dilakukan oleh pelaku seperti melemparkan lelucon tentang seks, memberikan komentar seks, menanyakan tentang kehidupan seks, dll. Dampak pelecehan seksual terhadap korban bermacam tergantung kadar dan durasi pelecehan yang dilakukan oleh pelaku.

Pelaku pelecehan seksual biasanya melakukan modus awal seolah pelaku ingin memiliki hubungan asmara dengan korban. Pada awalnya pelaku mendekati korban dan menunjukan perilaku yang dapat membuat korban memiliki citra positif terhadap pelaku. Setelah membuat korban merasa nyaman dan terbiasa dengan pelaku lalu ia akan melancarkan aksinya. Namun tidak semua pelaku melakukan hal tersebut, terdapat juga pelaku yang langsung melancarkan aksi.

C. Pembelajaran Mesin

Pembelajaran mesin merupakan salah satu cabang ilmu komputer di bidang kecerdasan buatan. Pembelajaran mesin

merupakan konsep yang berprinsip bahwa komputer atau mesin dapat melakukan sesuatu tanpa diprogram. Komputer dapat belajar dari data yang diberikan sehingga saat komputer mendapatkan data yang asing komputer bisa beradaptasi berdasarkan data tersebut. Algoritma untuk pembelajaran mesin sudah ada sejak lama namun yang membuat pembelajaran mesin berkembang saat ini yaitu perkembangan kecepatan komputasi dan Big Data.

Pembelajaran mesin sangat penting karena pembelajaran mesin dapat menghasilkan kesimpulan yang tepat. Pembelajaran mesin juga dapat memberikan model berdasarkan variasi data yang diberikan. Pembelajaran mesin sering digunakan perusahaan besar dalam menganalisis data. Hasil dari pembelajaran mesin digunakan sebagai keputusan bisnis agar dapat memberikan keuntungan lebih.

D. Pemrosesan Bahasa Alami

Pemrosesan/pengolahan bahasa alami adalah cabang ilmu komputer yaitu pembelajaran mesin agar komputer dapat memahami, menafsirkan, memanipulasi bahasa manusia. Pemrosesan bahasa alami merupakan teknik yang sudah lama ada namun berkembang pesat saat ini mengikuti perkembangan komputer yang memiliki kemampuan komputasi yang lebih cepat dan juga *big data* yang dapat menyimpan banyak kosa kata manusia.

Interaksi antara komputer dengan manusia agar dapat melakukan pemrosesan bahasa alami bisa beruba banyak hal misalnya komputer menangkap data berupa teks, komputer meminta perintah suara dari pengguna, komputer mengubah audio menjadi teks, dll. Saat ini pemrosesan bahasa alami mulai banyak diimplementasikan seperti pada Google Translate, Siri, Cortana, perangkat lunak pemeriksa tata bahasa, dll. Pemrosesan bahasa alami dapat dimanfaatkan di berbagai sektor kehidupan.

E. Rapid Automatic Keyword Extraction

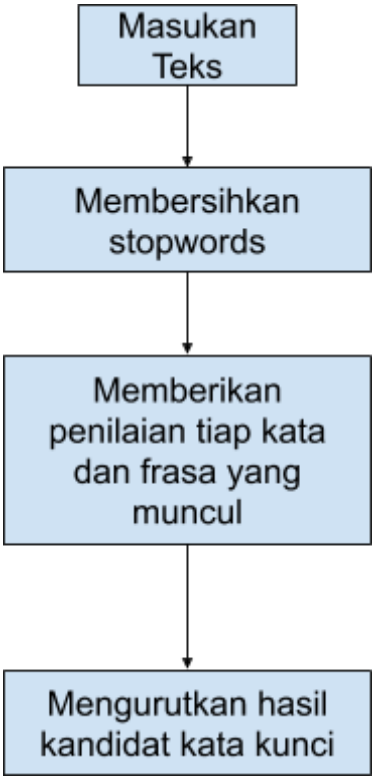
Salah satu algoritma penting dari pemrosesan bahasa alami adalah ekstraksi kata kunci. Banyak algoritma untuk ekstraksi kata kunci seperti *Text Frequency Inverse Document Frequency(TF-IDF), cosine similarity, text ranking,* dll. Namun yang ingin digunakan saat ini adalah *rapid automatic keyword extraction.* Algoritma ini memiliki kemampuan untuk mengekstraksi kata kunci penting dari suatu teks. Algoritma ini memanfaatkan teori probabilitas dan statistika pada kemunculan kata atau kombinasi kata. *Rapid automatic keyword extraction* memberikan hasil yang cukup presisi terhadap keyword yang dihasilkan. Selain itu algoritma ini juga memiliki kecepatan pemrosesan yang lebih cepat dibandingkan melakukan *text ranking.*

Algoritma ini memulai ekstraksi kata kunci dengan mengeliminasi *stopwords* yang berada di dalam teks. *Stopwords* adalah kata yang dianggap tidak memiliki arti penting dalam kalimat sehingga hilangnya *stopwords* tidak memengaruhi maksud dari kalimat tersebut. *Stopwords* bisa terdiri dari kata-kata yang sering digunakan dalam suatu bahasa atau kata-kata yang memang sengaja dianggap keberadaanya tidak memiliki arti penting dalam suatu program dan bila kata tersebut dihapus tidak akan merubah makna yang dimiliki oleh teks tersebut.

Setelah dilakukan eliminasi kata-kata yang dianggap tidak memengaruhi makna dari suatu kalimat atau teks, selanjutnya dilakukan penilaian suatu frasa berdasarkan *co-occurrence* atau frekuensi kemunculan kata secara berdampingan pada suatu teks. Penilaian ini disimpan dalam sebuah matriks untuk kemudian dilakukan penilaian kata yang dianggap penting secara individu dengan dilakukan perbandingan jumlah kata tersebut muncul ditambah dengan jumlah kata berdampingan yang muncul dengan kata tersebut dan jumlah kata tersebut muncul.

$$Penilaian = \frac{frekuensi\ kata + jumlah\ kata\ yang\ berdampingan}{frekuensi\ kata}$$

Lalu selanjutnya melakukan penilaian terhadap frasa yang muncul lebih dari sekali di dalam teks. Kata dan frasa yang telah dilakukan penilaian merupakan kandidat kata kunci yang akan dihasilkan. Langkah terakhir dilakukan pembatasan kuota kata kunci dengan urutan hasil penilaian terhadap kandidat kata kunci tersebut.



Gambar 2. Algoritma kasar *rapid automatic keyword extraction*
sumber: dokumen penulis

F. Regular Expression

Regular expression adalah salah satu metode dalam pencarian data dalam string. *Regular expression* tidak seperti algoritma pencocokan string pada umumnya karena didalam *regular expression* terdapat dua tipe karakter yaitu *literal character* dan *metacharacter. Literal character* ialah karakter biasa yang benar-benar ada wujudnya misalnya seperti yang terdapat pada ASCII yang terdiri dari berbagai huruf, angka,

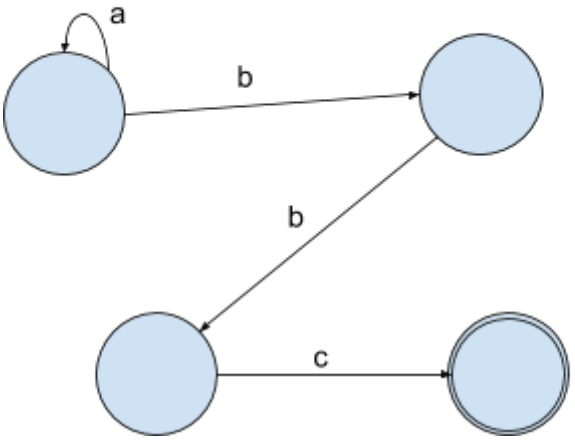
hingga tanda baca. Lalu yang membuat *regular expression* ini lebih unggul dari algoritma pencocokan string lainnya yaitu adanya *metacharacter. Metacharacter* adalah karakter yang tidak memiliki wujud seperti *literal character* namun memiliki aturan yang bisa menunjukan apakah suatu string dapat diterima atau tidak. Singkatnya *metacharacter* seperti pola atau ciri sebagai acuan dalam pencarian string. *Metacharacter* pada umumnya ditunjukan seperti pada tabel berikut:

Metacharacter	Artinya
.	Sesuai dengan satu karakter apapun
[...]	Sesuai dengan karakter yang terdapat dalam kurung siku
[^...]	Sesuai dengan karakter yang tidak terdapat dalam kurung siku
?	Boleh dicocokan namun pencocokan bersifat tidak wajib, bila tidak ada bisa dilanjut
*	Suatu karakter dengan jumlah pengulangan sebanyak apapun mulai dari nol
+	Suatu karakter dengan jumlah pengulangan sebanyak apapun mulai dari satu
|	Memberikan pilihan pola *regular expression*
^	String yang diawali dengan string sesudah karakter tersebut
$	String yang diakhiri dengan string sesudah karakter tersebut
{X,Y}	karakter dengan jumlah X-Y, Y tidak wajib jika hanya diisi X menjadi karakter dengan perulangan sejumlah X

Tabel 1. *Metacharacter regular expression*

Metacharacter pada *regular expression* menghasilkan *finite state machine (*FSM*)* untuk mengolah string masukan. FSM adalah alur dalam memeriksa masukan berdasarkan kondisi status yang ada. FSM memiliki luaran diterima atau tidak. Masukan diterima apabila kondisi status terakhir sesuai dengan status akhir FSM.

Regular expression = a*b{2}c



Gambar 3. Contoh *regular expression,* sumber: dokumen penulis

III. IMPLEMENTASI

Pencegahan pelecehan seksual terjadi di media sosial dapat dilakukan dengan mendeteksi kata kunci yang memiliki makna pelecehan seksual pada pesan yang akan dikirim. Kata kunci yang digunakan sebagai standar pesan mengandung pelecehan seksual didapatkan dengan menerapkan algoritma ekstraksi kata kunci. Algoritma *rapid automatic keyword extraction* membutuhkan dua masukan yaitu teks yang ingin diekstraksi dan *stopwords* atau kata kunci yang akan dieliminasi.

A. Memilih Acuan Standar Pelecehan Seksual

Acuan kata kunci pelecehan seksual berbeda-beda berdasarkan gaya bahasa yang digunakan suatu daerah. Kata kunci yang digunakan diekstraksi dari riwayat teks percakapan yang memang dianggap sebagai pelecehan seksual. Percakapan tersebut kemudian dapat diekstraksi menjadi beberapa kata kunci yang dapat digunakan untuk mendeteksi ujaran pelecehan seksual pada pesan yang akan dikirim di media sosial.

Riwayat teks percakapan dipilih karena teks percakapan lebih mudah diolah dibandingkan mengolah media seperti gambar atau video. Pelaku menggiring opini target juga menggunakan pesan-pesan yang dapat meluluhkan hati dan membuat target merasa nyaman. Secara logika pelaku tidak

mungkin melakukan pelecehan seksual secara langsung karena itu sia-sia.

B. Memilih *Stopwords*

Mungkin kita bisa saja dengan mudah mengetahui ujaran tidak senonoh yang mengarah ke pelecehan seksual. Akan tetapi mesin tidak bisa melakukannya. Mesin membutuhkan data pembanding untuk menentukan apa yang mesin cari. Pemilihan *stopwords* yang tepat memberikan hasil yang sesuai dengan yang diinginkan.

Pelecehan seksual erat hubungannya dengan hubungan asmara. Percakapan yang mengandung pelecehan seksual mirip dengan percakapan dua orang yang sedang menjalin hubungan asmara hanya saja ditambahkan dengan unsur seksual. Memilih percakapan orang yang sedang menjalin hubungan asmara sebagai *stopwords* dapat meningkatkan presisi kata kunci yang dihasilkan untuk mendeteksi pelecehan seksual.

C. Mengekstraksi Kata Kunci

Percakapan yang terbukti terindikasi pelecehan seksual diekstraksi agar didapatkan kata kunci. Proses ekstraksi dilakukan menggunakan algoritma *rapid automatic keyword extraction* dengan menggunakan *stopwords* yang telah dipilih sebelumnya. Berikut kode program implementasi ekstraksi kata kunci pelecehan seksual menggunakan bahasa python dan algoritma *rapid automatic keyword extraction* menggunakan pustaka 'muti-rake'.

ekstrak.py

```
import re
from multi_rake import stopwords,Rake
from pyrsistent import thaw

def ekstrak(percakapanPelecehan,
pacaranBiasa):
	#Mengekstraksi kata kunci yang dapan
menunjukan pelecehan seksual

	#stopwords umum pada bahasa indonesia
	stopwordsID =
thaw(stopwords.STOPWORDS.get('id'))
	#stopwords yang dihasilkan dari
percakapan pacaran
	removedSign = re.sub('[^A-Za-z]',' 
',pacaranBiasa)
	stopwordPacaran =
set(removedSign.split(' '))

	#Stopword hasil gabungan kedua
stopword
	useStopword = stopwordsID |
stopwordPacaran
	#kakas rapid automatic keyword
extraction
	rake=Rake(stopwords=useStopword)
	return
rake.apply(percakapanPelecehan)
```

D. Mendeteksi Unsur Pelecehan Seksual

Setelah kata kunci berhasil didapatkan selanjutnya kata kunci tersebut digunakan sebagai acuan terjadi pelecehan seksual. Pesan teks yang akan dikirim pada media sosial akan diperiksa apakah mengandung kata kunci yang telah dihasilkan sebelumnya. Jika pesan yang akan dikirimkan mengandung salah satu dari kata kunci maka pesan tersebut merupakan pesan yang mengandung unsur pelecehan seksual sehingga pesan tersebut bisa dicekal agar korban dapat terlindungi dari pelecehan seksual.

Deteksi unsur pelecehan seksual diimplementasikan menggunakan pustaka *regular expression* pada bahasa python. Algoritma untuk mendeteksi unsur pelecehan seksual menggunakan algoritma untuk melakukan pencarian biasa. Jika saat pencarian telah menemukan salah satu kata kunci pada pesan teks maka program akan memberikan nilai *true.* Berikut kode program untuk mendeteksi unsur pelecehan seksual.

deteksi.py

```
import re

def deteksi(pesanTeks,keywords):
	i = 0
	while i<len(keywords) and not
re.search(keywords[i],pesanTeks):
	i+=1

	return i<len(keywords)
```

IV. Pengujian

Pengujian dilakukan menggunakan sistem operasi ubuntu menggunakan python versi 3.6.9. Data untuk pengujian penulis mengarang namun karangan tersebut telah diusahakan mendekati dengan data kasus pelecehan seksual sesungguhnya sesungguhnya.

Data buatan untuk diekstrak menjadi kata kunci yang merepresentasikan pelecehan seksual. Data yang dimasukan merupakan pendekatan yang dipikirkan penulis bagaimana bentuk pelecehan seksual pada percakapan teks.

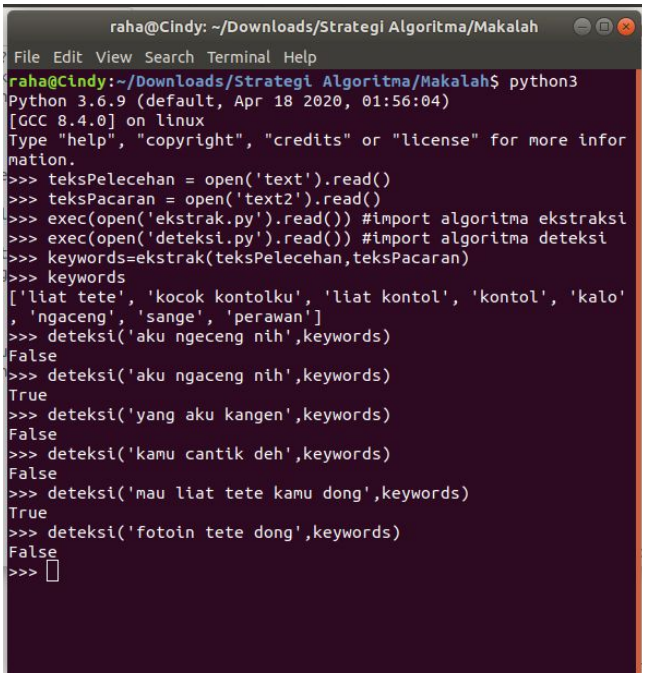
```
kamu lagi ngapain?
lagi tiduran aja kok, kalo kamu?
```

aku lagi makan nih
makan apa?
makan kamu
apaan sih
kamu udah makan belom?
belom
kamu udah pernah liat kontol?
...
kamu mau liat kontol aku?
kontol aku ngaceng nih
yang aku mau liat tete kamu dong
kamu cantik deh
aku sange
aku kocok kontolku yaa
kamu masih perawan yaa

Data buatan yang akan dijadikan *stopwords* untuk proses ekstraksi percakapan pelecehan seksual. Data merupakan pendekatan penulis tentang percakapan orang yang menjalin hubungan asmara.

kamu lagi ngapain?
lagi tiduran aja kok
yang aku lapar, ayo makan keluar
wah iya ayo aku juga lapar
makan dimana?
terserah kamu
kamu cantik deh tapi belom mandi
aku udah mandi yaa dasar
pap dong aku kangen
nih
tuh kan belom mandi tapi tetap cantik
ayo makan keluar cepet
oke aku jemput yaa
udah sampe rumah?
udah
kok belum tidur?
kamu chat sih
semangat menjalani hari ya

Hasil eksekusi program:



Gambar 4. Hasil uji coba, sumber: dokumen penulis

Dari percobaan dilakukan pemeriksaan terhadap empat pesan dan hasilnya sebagai berikut:

Pesan	Terdeteksi	Sesuai Prediksi
aku ngaceng nih	✓	✓
yang aku kangen		✓
kamu cantik deh		✓
fotoin tete dong		

Tabel 2. Hasil uji coba.

Dari empat percobaan yang dilakukan terlihat bahwa 3 percobaan berhasil. Percobaan gagal terjadi karena metode pendeteksi ujaran pelecehan seksual ini menggunakan pembelajaran mesin. Mesin belajar dari data yang disediakan dan data yang dibuat penulis mungkin tidak lengkap atau mungkin tidak mencakup sekian banyak kemungkinan skenario. Hal ini dapat dioptimasi dengan memberikan data lebih sehingga pendeteksi bisa lebih presisi.

V. Kesimpulan

Media sosial bisa saja menjadi tempat kejahatan terjadi. Hal tersebut terjadi karena masih ada celah kejahatan yang bisa dilakukan. Bang Napi pernah berkata “Kejahatan terjadi bukan hanya karena ada niat pelakunya, tetapi juga karena adanya kesempatan. WASPADALAH! WASPADALAH!”. Pada media sosial kesempatan untuk melakukan kejahatan sangat banyak namun semua seharusnya bisa diatasi dengan kemajuan teknologi.

Perkembangan komputer dalam memahami bahasa alami manusia penulis manfaatkan untuk melakukan pencegahan terhadap pelecehan seksual pada media sosial. Metode yang penulis gunakan tidak hanya bisa digunakan hanya pada pelecehan seksual saja.

Selain digunakan sebagai keperluan bisnis, pembelajaran mesin juga dapat digunakan pada sektor sosial masyarakat. Mungkin jika metode diterapkan dapat mengurangi kasus pelecehan seksual pada media sosial.

Tautan Video pada Youtube

https://youtu.be/2py8iqAvJxo

Ucapan Terima Kasih

Pertama penulis ingin mengucapkan terimakasih kepada Tuhan Y.M.E. atas berkat dan hidayah-Nya penulis bisa menyelesaikan makalah ini. Terima kasih kepada Dr. Masayu Leylia Khodra, Dr. Nur Ulfa Maulidevi, S.t, M.Sc., dan Dr. Ir. Rinaldi Munir, M.T. sebagai dosen pengampu mata kuliah Strategi Algoritma yang telah memberikan ilmunya. Selain itu terima kasih juga kepada orang tua dan teman-teman yang telah mendukung dalam penulisan makalah ini. Semoga makalah ini dapat berkontribusi dalam perkembangan teknologi dan dapat bermanfaat pada masyarakat.

References

[1] Endah Trijiwati, "Pelecehan Seksual: Tinjauan Psikologi," http://journal.unair.ac.id/filerPDF/Pelecehan%20Seksual%20Tinjauan%20Psikologi.pdf, diakses 26 April 2020 Pukul 02.02.

[2] "Apa itu pemrosesan bahasa alami?", https://www.sas.com/id_id/insights/analytics/what-is-natural-language-processing-nlp.html, diakses 2 Mei 2020 pukul 01.27

[3] Rose, Stuart & Engel, Dave & Cramer, Nick & Cowley, Wendy. (2010). Automatic Keyword Extraction from Individual Documents. 10.1002/9780470689646.ch1.

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.

Gresik, 2 Mei 2020

M Rahadian Alamsyah P W
13518011