Algoritma Pencocokan Konten Artikel Dengan Menerapkan *Regular Expression* Untuk Mendeteksi Plagiarisme

Naufal Zhafran Latif
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
latifnaufal7@gmail.com

***Abstract*—Artikel merupakan suatu media atau wadah bertukar informasi melalu tulisan. Konten merupakan aspek paling penting dari sebuah artikel. Konten tentu harus menarik dan berbobot. Sebuah artikel akan menjadi lebih baik jika memiliki konten yang baru dan tidak melakukan plagiariesme terhadap artikel lain. Dengan menggunakan algoritma *regular expression* yang dimodifikasi, penulis artikel dapat melihat tingkat kecocokan sebuah artikel dengan artikel lain berdasarkan konten yang dibawa.**

Keywords—artikel;konten;regularexpressin;plagiarisme

I. PENDAHULUAN

Artikel merupakan salah satu bentuk tulisan yang sering ditulis oleh manusia. Sebuah artikel menjelaskan tentang sebuah topik yang dibawakan oleh penulisnya. Media untuk melalukan penulisan dan penyebaran sebuah artikel sangat beragam. Bisa melalui media cetak,media massa maupun media elektronik.

Artikel menjadi wadah manusia untuk mengeskpresikan dirinya. Semua pemikiran, perasaan dan ilmu yang ada pada diri seorang penulis maupun orang-orang pada umumnya dapat dituangkan dalam bentuk sebuah artikel. Artikel pun bisa menjadi sumber ilmu pengetahuan untuk orang-orang yang membaca artikel tersebut. Selain itu banyak penulis artikel yang cukup kompeten dalam melakukan penulisan menjadikan artikel tersebut sebuah produk yang dapat memberikan penulis keuntungan secara finansial. Tentu untuk sampai pada tingkat ini diperlukan ketekunan dan latihan dalam menulis sebuah artikel yang baik, menarik dan bernilai.

Sebuah artikel yang baik harus memiliki sistematika penulisan yang menarik, jelas dan membuat pembacanya tidak bosan. Selain itu artikel harus memiliki konten yang bagus, hangat dan unik. Syarat-syarat merupakan hal yang penting jika sebuah artikel ingin banyak dibaca orang lain karena artikel yang buruk akan membuat pembaca bosan dan tidak membaca artikel yang penulis tersebut buat. Salah satu hal yang sulit dari membuat artikel yang baik adalah sulitnya menemukan sebuah konten atau ide untuk menulis artikel yang unik karena ide yang unik dan orisinil adalah hal yang cukup langka.

Ide yang unik dan belum ada yang menulis merupakan suatu hal yang cukup langka. Untuk mendapatkan ide atau konten yang seperti itu diperlukan pengetahuan yang luas dan pengalaman yang banyak. Tidak semua penulis terutama penulis yang masih baru dapat memiliki pengalaman dapat dengan mudah menemukan ide konten yang orisinil dan unik.. Hambatan semacam ini memicu tumbuhnya plagiarisme terhadap konten sebuah artikel.

Plagiarisme yang terjadi bisa disebabkan dua hal. Pertama memang penulis artikel yang sengaja melakukan plagiarisme sehingga penulis tersebut tidak perlu berpikir lebih lama untuk menulis artikel tersebut. Hal ini memang hanya dapat dihindari dengan kesadaran diri sendiri untuk tidak melakukan plagiarisme. Tetapi memang sudah ada cara lain seperti membuat landasan hukum mengenai plagiarisme itu sendiri walaupun hukum tersebut bersifat umum.

Kedua penulis tersebut tidak tahu bahwa artikel yang ditulisnya memiliki konten yang relatif sama dengan artikel lain. Hal ini terkadang sulit untuk dihindari karena sebagai penulis belum tentu punya waktu untuk membaca semua artikel dengan konten yang mirip. Selain itu terkadang konten yang dijelaskan dalam sebuah artikel belum tentu sama persis secara kata-kata dengan apa yang dibawakan oleh penulis artikel. Seperti konten yang sebenarnya adalah parafrase dari konten yang berusaha penulis artikel itu tulis.

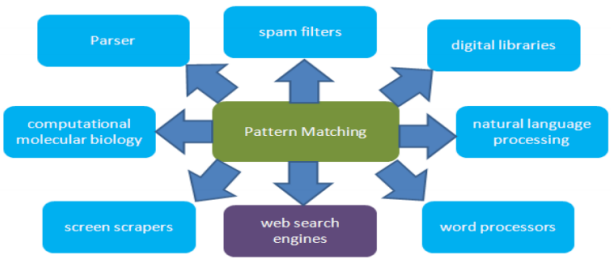
Untuk itu diperlukan sebuah alat yang dapat mempermudah penulis artikel dengan melakukan pengukuran terhadap artikel-artikel yang ada. Hal yang diukur adalah tingkat kecocokan kedua buah artikel dalam sisi konten yang dibawakan. Dengan alat ini penulis artikel akan mudah mendeteksi apakah tulisan yang berusaha dia tulis sudah pernah ditulis oleh orang lain.

II. LANDASAN TEORI

A. *Algoritma Pencocokan Pola*

Algoritma pencocokan pola merupakan sebuah algoritma yang mengecek sekumpulan urutan string memiliki sebuah pola sesuai dengan pola yang telah ditentukan diawal. Algoritma ini akan menghasilkan keberadaan pola yang dicari pada sekumpulan string. Biasanya algoritma ini dipakai dalam melakukan pencarian pada sekumpulan string.

Kegunaan algoritma pencocokan pola sangat beragam. Algoritma ini dapat digunakan untuk melakukan pencarian sebuah kata , informasi yang terkandung sampai tingkat kenegatifan dari sebuah tulisan atau artikel. Selain itu algoritma ini merupakan algoritma yang sering dipakai dalam situs pencarian website seperti google.



Gambar 2.1 fungsi pattern matching[1]

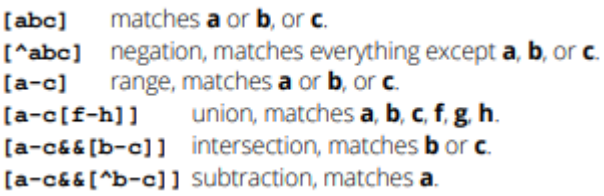
Algoritma ini bermacam-macam. Ada yang melakukan pengecekan pola yang harus sama dengan pola yang diinginkan seperti *Knutt-Moris-Pratt Algorithm* dan *Boyer-Moore Algorithm*. Ada juga yang dapat melakukan pencarian pola dengan *query* pola yang kita inginkan seperti *Regular Expression*.

B. *Regular Expression*

Regular Expression atau yang biasa disingkat regex merupakan sebuah algoritma dan alat untuk melakukan pencarian pola yang cocok pada sebuah atau lebih karakter pada string. Kelebihan dari *Regular Expression* adalah pencocokan yang dilakukan tidak hanya pada sebuah pola yang statis dan spesifik, dia dapat mencocokan berbagai pola yang mirip ataupun satu himpunan kata atau karakter.

Pada awalnya *Regular Expression* hanya dipergunakan pada sistem operasi UNIX. Selama keberjalanan waktu, *Regular Expression* akhirnya dapat diimplementasikan dan dipergunakan pada banyak bahasa pemrograman. *Regular Expression* juga sering dipergunakan pada sebuah *Text Editor* seperti *Visual Studio Code* ataupun *Word Processing Application* seperti *Microsoft Word*.

Dalam sebuah bahasa pemrograman dan komputer, implementasi algoritma *Regular Expression* biasanya menggunakan *Finite Automata*. *Finite automata* atau *finite state machine* merupakan abstraksi mesin pada perangkat lunak sebuah komputer yang memiliki *state* dan transisi antara *state* tersebut. Tetapi kebanyakan bahasa pemrograman sudah menyiapkan *library* atau fungsi bawaan yang dapat melakukan pencocokan menggunakan *regular expression*.



Gambar 2.2 contoh syntax regular expression

C. Artikel

Artikel menurut Kamus Besar Bahasa Indonesia atau KBBI merupakan karya tulis lengkap, misalnya laporan berita atau esai dalam majalah, surat kabar dan sebagainya. Tujuan umum dari sebuah artikel dibuat adalah untuk mendidik, meyakinkan, memberitahu hingga menghibur pembacanya dan dipublikasikan melalui media cetak atau pun media *online*.

Artikel merupakan satu dari beberapa jenis tulisan yang ada dalam bahasa Indonesia. Artikel memiliki beberapa ciri-ciri yaitu :

1. Isi dari artikel tersebut harus bersifat fakta atau harus didapatkan dari sumber yang terpercaya. Penulis artikel harus dapat mempertanggung jawabkan tulisannya.
2. Metode penulisan sebuah artikel harus tepat dan sistematis. Hal ini dimaksudkan untuk mempermudah pembaca dalam memperoleh informasi.
3. Tulisan bersifat informatif dan benar adanya.Isi dari tulisan harus sesuai dengan fakta yang ada.
4. Tulisan harus dibuat dengan bahasa yang formal. Hal ini dimaksud agar masyarakat umum mudah dalam memahami maksud dari artikel.
5. Tulisan sebuah artikel tidak hanya berasa dari fakta-fakta yang ada. Tetapi ada unsur opini pribadi penulis yang tentu harus berdasarkan fakta yang benar.
6. Kalimat yang digunakan harus lugas, logis dan efektif.

Selain ciri-ciri tersebut, artikel memiliki tujuan yang spesifik mengenai alasan sebuah artikel dibuat yaitu:

1. Mendeskripsikan, menjelaskan, dan menguraikan suatu pokok pembahasan masalah yang sudah ditentukan
2. Mendeskripsikan suatu batasan kajian. Hal ini merupakan tujuan utama dari artikel ilmiah

Jika dilihat dari isi atau konten dari sebuah artikel, bisa dikategorikan menjadi lima macam yaitu:

1. Artikel Narasi

Artikel ini memuat runtutan dan cerita dari suatu kejadian yang telah terjadi. Cerita tersebut disajikan secara jelas dan informatif

2. Artikel Deskripsi

Artikel ini berisi gambaran mengenai suatu objek sehingga pembaca dapat seolah-oleh melihat, mendengar dan merasakan objek tersebut.

3. Artikel Argumentasi

Artikel ini memiliki tujuan khusus yaitu untuk membuktikan kebenaran dan keabsahan sebuah pendapat atau pemikiran dengan data dan fakta yang dapat dipertanggungjawabkan sebagai alasan dan bukti.

4. Artikel Persuasif

Artikel ini akan berusaha membuat pembacanya untuk terpengaruh terhadap konten pada artikel. Biasanya berupa ajakan atau seruan akan suatu tindakan.

5. Artikel Eksposisi

Artikel ini berisi uraian atau penjelasan tentang suatu topik dengan tujuan untuk memberikan pengetahuan dan informasi baru kepada pembancanya.

D. Plagiarisme

1. Definisi Plagiat

Berdasarkan Peraturan Mentri Pendidikan RI Nomor 17 Tahun 2017 bahwa Plagiat adalah perbuatan sengaja atau tidak sengaja dalam memperoleh atau mencoba memperoleh kredit atau nilai untuk suatu karya ilmiah, dengan mengutip sebagian atau seluruh karya dan atau karya ilmiah pihak lain yang diakui sebagai karya ilmiahnya, tanpa menyatakan sumber secara tepat dan memadai.

Jika dilihat dari Kamus Besar Bahasa Indonesia atau KBBI plagiat adalah pengambilan karangan (pendapat dan sebagainya) orang lain dan menjadikannya seolah-olah karangan (pendapat) sendiri.

2. Ruang Lingkup Plagiarisme

Berdasarkan kedua definisi diatas dapat ditarik kesimpulan bahwa ruang lingkup plagiarisme yaitu :

a. Mengutip kata-kata atau kalimat orang lain tanpa menggunakan tanda kutip dan tanpa menyebutkan identitas sumbernya.
b. Menggunakan gagasan, pandangan atau teori orang lain tanpa menyebutkan identitas sumbernya.
c. Menggunakan fakta (data, informasi) milik orang lain tanpa menyebutkan identitas sumbernya.
d. Mengakui tulisan orang lain sebagai tulisan sendiri.
e. Melakukan parafrase (mengubah kalimat orang lain ke dalam susunan kalimat sendiri tanpa mengubah idenya) tanpa menyebutkan identitas sumbernya.
f. Menyerahkan suatu karya ilmiah yang dihasilkan dan /atau telah dipublikasikan oleh pihak lain seolah-olah sebagai karya sendiri.

3. *Tipe Plagiarisme*

Plagiarisme memiliki beberapa tipe. Menurut Soelistyo[6] tipe-tipe tersebut yaitu :

a. Plagiarisme Kata demi Kata (Word for word Plagiarism). Penulis menggunakan kata-kata penulis lain (persis) tanpa menyebutkan sumbernya.
b. Plagiarisme atas sumber (Plagiarism of Source). Penulis menggunakan gagasan orang lain tanpa memberikan pengakuan yang cukup (tanpa menyebutkan sumbernya secara jelas).
c. Plagiarisme Kepengarangan (Plagiarism of Authorship). Penulis mengakui sebagai pengarang karya tulis karya orang lain.
d. Self Plagiarism. Termasuk dalam tipe ini adalah penulis mempublikasikan satu artikel pada lebih dari satu redaksi publikasi. Dan mendaur ulang karya tulis/ karya ilmiah. Yang penting dalam self plagiarism adalah bahwa ketika mengambil karya sendiri, maka ciptaan karya baru yang dihasilkan harus memiliki perubahan yang berarti. Artinya Karya lama merupakan bagian kecil dari karya baru yang dihasilkan. Sehingga pembaca akan memperoleh hal baru, yang benar-benar penulis tuangkan pada karya tulis yang menggunakan karya lama.

4. *Sanksi Untuk Pelaku Plagiarisme*

Undang-Undang No. 20 Tahun 2003 mengatur sanksi bagi orang yang melakukan plagiat, khususnya yang terjadi dilingkungan akademik. Sanksi tersebut adalah sebagai berikut (Pasal 70):

Lulusan yang karya ilmiah yang digunakannya untuk mendapatkan gelar akademik, profesi, atau vokasi sebagaimana dimaksud dalam Pasal 25 Ayat (2) terbukti merupakan jiplakan dipidana dengan pidana penjara paling lama dua tahun dan/atau pidana denda paling banyak Rp 200.000.000,00 (dua ratus juta rupiah).

Peraturan Menteri Nomor 17 Tahun 2010 telah mengatur sanksi bagi mahasiswa yang melakukan tindakan plagiat. Jika terbukti melakukan plagiasi maka seorang mahasiswa akan memperoleh sanksi sebagai berikut:

1. Teguran
2. Peringatan tertulis
3. Penundaan pemberian sebagian hak mahasiswa
4. Pembatalan nilai
5. Pemberhentian dengan hormat dari status sebagai mahasiswa
6. Pemberhentian tidak dengan hormat dari status sebagai mahasiswa
7. Pembatalan ijazah apabila telah lulus dari proses pendidikan.

III. IMPLEMENTASI DAN PEMBAHASAN

Dalam implementasi algoritma pencocokan konten artikel, penggunaan *Regular Expression* hanya bagian dari keseluruhan algoritma. Sebelum sebuah artikel bisa dicocokan dengan artikel lain harus dilakukan ekstraksi informasi pada artikel tersebut. Setelah dilakukan ekstraksi , data yang didapatkan baru bisa dicocokan menggunakan *Regular Expression*. Nanti akan dihasilkan tingkat kecocokan dari artikel yang diuji.

A. Ekstraksi Informasi

Esktraksi data merupakan langkah yang sangat penting pada algoritma pencocokan konten artikel ini. Data atau informasi dari artikel tidak bisa langsung digunakan untuk melakukan pencocokan. Selain itu banyak informasi yang tidak relevan atau tidak diperlukan dalam pencocokan konten artikel yang harus disaring dan dibuang.Hal ini dilakukan untuk mendapatkan hasil yang lebih optimal dan akurat.

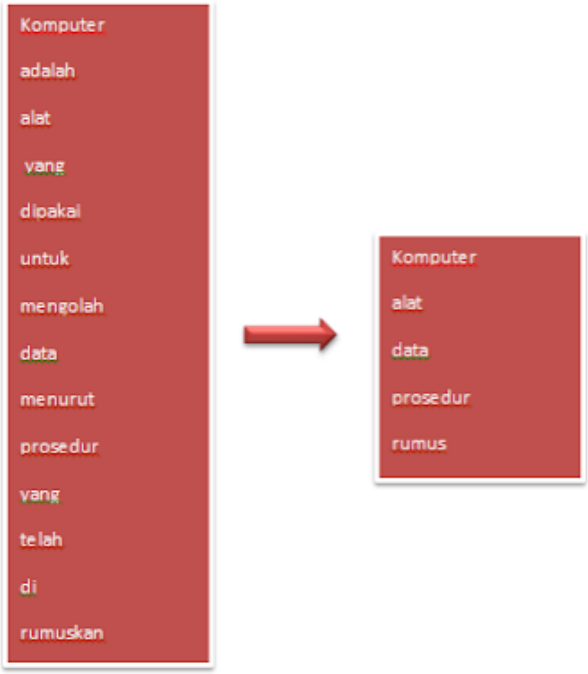
Sebelum melakukan ekstraksi data, artikel yang masih menjadi sebuah string panjang harus dipecah menjadi kata-kata. Pada algoritma ini sebuah kata di definisikan sebagai rangkaian kata-kata yg dibatasi oleh *whitespace* atau spasi. Itu artinya untuk kata “apapun” akan menjadi satu kata dan “apa pun” akan menjadi dua buah kata.

```
/* Variabel Data untuk menghitung
kemunculan suatu kata*/

For kata in artikel1:
  Data[kata] = Data[kata] + 1
```

Untuk mendapatkan informasi yang diperlukan dalam pencocokan konten artikel, *keyword* atau kata kunci dari sebuah artikel. Cara yang digunakan pada algoritma ini adalah dengan menghitung frekuensi atau jumlah kemunculan semua kata pada artikel tersebut. Kita bisa mengasumsikan 25% dengan jumlah kata terbanyak dari seluruh jenis kata adalah kata kunci dari artikel itu.

Akan tetapi ada kondisi dimana kata-kata yang sebenarnya tidak penting dapat terpilih sebagai sebuah kata kunci. Kata-kata yang dimaksud seperti dan,atau,yang, dan kata-kata sambung lainnya. Kata-kata semacam ini dapat memiliki frekuensi kemunculan yang tinggi karena sering digunakan untuk menyambungkan kata dengan kata maupun kalimat dengan kalimat. Untuk itu kata-kata seperti ini dapat dieliminasi dari daftar jumlah kata yang kita dapat dari langkah sebelumnya. Sehingga akan didapatkan daftar kata kunci yang lebih akurat.



Gambar 3.1 Contoh Penyaringan Stop Words pada String[5]

Setelah didapatkan kata-kata kunci tersebut, baru dapat dilakukan pencocokan terhadap artikel lain. Pencocokan kali ini dilakukan dengan algoritma *Regular Expression*. Tetapi sebelum dilakukan pencocokan, artikel yang akan dipakai untuk melakukan pencocokan juga harus dilakukan penyaringan informasi agar hasil lebih akurat. Sama seperti sebelumnya, kata-kata yang tidak penting pada artikel akan dieliminasi sehingga hanya tersisa kata-kata yang diasumsikan penting.

```
/*stopwords merukapan basis data
stopwords atau kata yang tidak
penting*/

For kata in artikel1:
  For stopword in stopwords:
    If(kata != stopword):
       Data[kata] = Data[kata] + 1

KataKunci = Data[25% dari total data
dengan jumlah terbanyak]
```

B. Pencocokan Menggunakan Regular Expression

Setelah artikel yang akan dicocokan telah dihilangkan kata yang tidak penting. Pencocokan dengan *Regular Expression* dapat dilakukan. Pertama artikel akan dihitung total kata yang tersisa setelah dihilangkan kata yang tidak penting. Lalu

dilakukan pencocokan dengan menggunakan *Regular Expression* untuk setiap kata kunci yang sudah kita dapatkan tadi.

Akan dihitung jumlah kemunculan kata kunci tersebut pada artikel. Setelah itu jumlah kemunculan kata kunci pada artikel akan digabungkan dan dihitung totalnya. Dari total tersebut dibagi dengan total kata dari artikel terakhir akan dihasilkan persentase kemiripan konten dari kedua artikel yang diuji.

```
/* Katakunci didapatkan dari langkah yang
sebelumnya*/

For Kata in Katakunci:
  For Kataartikel in artikel2:
    If(Kata == Kataartikel):
       NKata[Kata] = NKata[Kata] + 1

/*Sum merupakan fungsi yang menjumlahkan
jumlah dari semua kata*/

/*Length merupakan fungsi untuk menghitung
jumlah kata pada artikel*/

Persentase = Sum(NKata) / Lenght(artikel2)
```

C. Pencocokan Dengan Melihat Sinonim dari Kata Kunci

Kata kunci dari sebuah artikel cukup untuk merepresentasikan topik atau konten yang sedang dibawa oleh artikel tersebut. Tetapi dalam penulisan artikel terkadang penulis memparafrase kalimat-kalimat pada artikel. Parafrase ini salah satunya dengan mengubah kata menjadi sinonim dari kata tersebut sehingga tidak terlihat sama.

Mendeteksi sebuah sinonim dari kata diperlukan sebuah basis data sinonim yang besar. Hal ini dikarenakan sebuah kata dapat memiliki lebih dari satu sinonim. Basis data sinonim ini untuk bahasa indonesia bisa berdasarkan kamus besar bahasa indonesia.

Pencocokan sinonim ini akan menjadi kelanjutan setelah algoritma ini melakukan pencocokan terhadap kata kunci. Proses pencocokan tetap dilakukan menggunakan algoritma yang sama dengan pencocokan kata kunci.

```
/* Katakunci didapatkan dari langkah yang
sebelumnya*/

/* isSinonim merupakan fungsi yang
terhubung dengan basis data sinonim. Fungsi
ini berguna untuk melakukan pengecekan
apakah kedua kata tersebut termasuk
sinonim*/

For Kata in Katakunci:
  For Kataartikel in artikel2:
    If(Kata == Kataartikel &&
isSinonim(Kata,KataArtikel):
       NKata[Kata] = NKata[Kata] + 1

/*Sum merupakan fungsi yang menjumlahkan
jumlah dari semua kata*/

/*Length merupakan fungsi untuk menghitung
jumlah kata pada artikel*/

Persentase = Sum(NKata) / Lenght(artikel2)
```

IV. KESIMPULAN DAN SARAN

Dengan memanfaatkan *Regular Expression* dan beberapa tahap untuk menyaring informasi dari sebuah artikel, dapat ditentukan tingkat kecocokan konten dari dua buah artikel. Tentu hal ini akan mempercepat proses untuk menentukan sebuah plagiarisme atau tidak karena tidak memerlukan proses membaca yang rinci secara manual. Akan tetapi algoritma ini tentu bukan algoritma yang sangat akurat. Bahasa bukan merupakan sebuah ilmu yang pasti. Banyak hal-hal dalam suatu bahasa yang tidak memiliki pola dan hanya bisa dimaknai oleh orang-orang yang memiliki insting.

Untuk pengembangan algoritma ini dapat menggunakan *Artificial Intellegence* atau kecerdasan buat agar pencocokan semakin akurat. Terutama dengan perkembangan *Machine Learning* sehingga algoritma ini dapat belajar dan meningkatkan keakuratan seiring keberjalanan waktu.

ACKNOWLEDGMENT

Penulis ingin mengucapkan terima kasih yang sebesarbesarnya kepada Tuhan Yang Maha Esa, berkat rahmat-Nya penulis dapat menyelesaikan makalah ini tepat waktu. Penulis juga ingin mengucapkan terimakasih kepada Bapak-Ibu Dosen Strategi Algoritma ITB yang telah membimbing penulis selama satu semester sehingga penulis mendapatkan banyak sekali manfaat dan pengetahuan dari mata kuliah ini, yaitu kepada Dr. Nur Ulfa Maulidevi, ST, M.Sc. Juga semua orang dan sumber yang telah membantu proses penyelesaian makalah ini sehingga penulis berhasil menyelesaikan makalah ini dengan tepat waktu.

REFERENCES

[1] Diwate, Rahul. (2013). Study of Different Algorithms for Pattern Matching.

[2] "Regular Expression." Regular Expression Definition, Apr. 2019, techterms.com/definition/regular_expression. diakses 26 april 2019

[3] "Implementing a Regular Expression Engine." · Denis Kyashif's Blog, 17 Feb. 2019, deniskyashif.com/implementing-a-regular-expression-engine/. diakses 26 april 2019

[4] "REGULAR EXPRESSIONS." Regular Expressions, Apr. 2019, users.cs.cf.ac.uk/Dave.Marshall/Internet/NEWS/regexp.html. diakses 26 april 2019

[5] "Text Mining." Text Mining, 1 Jan. 1970, johansyah-data-mining.blogspot.com/2017/09/text-mining.html. diakses 26 april 2019

[6] "Panduan Anti Plagiarism." Perpustakaan, 5 Sept. 2014, lib.ugm.ac.id/ind/?page_id=327. diakses 26 april 2019

[7] Zakky, Oleh. "Pengertian Artikel Beserta Penjelasan Dan Ciri-Ciri Artikel [Lengkap]." ZonaReferensi.com, 26 Aug. 2018, www.zonareferensi.com/pengertian-artikel/.. diakses 26 april 2019

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, 26 April 2019

Naufal Zhafran Latif 13517095