Analisis Sentimen Situs Pembajak Artikel Penelitian Menggunakan Metode *Lexicon-Based*

Michella G. Undap[1], Vivi P. Rantung[2], Parabelem T. D Rompas[3]

Program Studi Teknik Informatika UNIMA
michellaaundap@gmail.com

***Abstract*— The development of technology is very rapid in today's era, especially with the development of internet technology. In the field of education itself, publication and access to research references can be easily done via the internet. However, with the internet also, research journal piracy can occur when access to references cannot be obtained due to paywall barriers. Expressions of this can be found on social media, where users can use the platform to express their opinions freely. The availability of these data on the internet that can be retrieved and analyzed so as to produce information. In this study, we will analyze public sentiment towards one of the research article hijacking sites, namely Sci-Hub. The method used is Lexicon-based which is a method that uses a dictionary for classifying sentiment text. The stages of the method which are data collection, loading a dictionary, data pre-processing, keyword extraction to determining sentiment as the final stage to get sentiment classification results. The data used are 35,755 tweets with 28,193 English-language tweets and 7,552 Indonesian-language tweets. The data is classified according to the rules in the InSet dictionary with the results of the classification of positive sentiment 59.51%, negative sentiment 26.42% and neutral sentiment 14.07%.**

.

***Keywords*—sentiment analysis, lexicon-based, sci-hub, twitter, tweet, python.**

***Abstrak*— Perkembangan teknologi sangatlah pesat pada era sekarang ini, apalagi dengan perkembangan teknologi internet. Di bidang pendidikan sendiri, untuk publikasi serta akses referensi penelitian bisa mudah dilakukan dengan internet. Namun dengan internet pula, pembajakan jurnal penelitian bisa terjadi bilamana akses untuk referensi tidak bisa diperoleh karena hambatan *paywall*. Ekspresi terhadap hal tersebut bisa didapati di sosial media, dimana penggunanya bisa menggunakan *platform* tersebut untuk menuangkan opini mereka secara bebas. Ketersediaan data tersebut di internet yang dapat diambil serta dianalisis sehingga menghasilkan suatu informasi. Pada penelitian ini akan menganalisis sentimen masyarakat terhadap salah satu situs pembajak artikel penelitian yaitu Sci-Hub. Metode yang digunakan yaitu *Lexicon-based* yang merupakan metode yang memanfaatkan *dictionary* untuk pengklasifikasian teks sentimen. Tahapan metode yang dilakukan yaitu, pengambilan data, memuat kamus, *pre-processing* data, ekstraksi kata kunci hingga penentuan sentimen sebagai tahapan akhir untuk mendapatkan hasil klasifikasi sentimen. Data yang digunakan adalah *tweets* yang berjumlah 35.755 dengan *tweets* berbahasa inggris sebanyak 28.193 dan *tweets* berbahasa indonesia sebanyak 7.552. Data diklasifikasikan sesuai dengan *rule* dalam *dictionary* InSet dengan hasil klasifikasi sentimen positif 59.51%, sentimen negatif 26.42% dan sentimen netral 14.07%.**

***Kata Kunci*—analisis sentimen, *lexicon-based*, sci-hub, *twitter*, *tweet*, python.**

I. PENDAHULUAN

Perkembangan teknologi sekarang ini sangat membantu banyak hal dalam kehidupan masyarakat pada umumnya, contohnya teknologi internet yang menjadi sarana penyebaran informasi dari dan ke bagian seluruh dunia. Dalam bidang penelitian, beberapa manfaat teknologi internet yang sering didapati yaitu menjadi sarana dalam membantu publikasi serta penyedia akses untuk referensi penelitian. Hak akses artikel penelitian dari situs-situs penyedia publikasi tersebut ada yang tersedia secara gratis dan juga berbayar yang akhirnya memunculkan situs-situs pembajak, yang tujuannya agar artikel-artikel penelitian bisa diakses secara gratis. Banyak pro serta kontra dari masyarakat yang mereka tuangkan di media sosial mengenai keberadaan situs pembajak artikel penelitian ini.

Menentukan opini publik terhadap topik situs pembajak artikel penelitian dapat dilakukan dengan mengambil data dari media sosial. Salah satu media sosial yang sering digunakan masyarakat adalah *Twitter* yang merupakan salah satu platform *micro-blog* [1][2] yang adalah platform dalam bentuk *blog* yang pembaharuan teksnya singkat. Salah satu cara untuk memanfaatkan data *Twitter* dalam penelitian adalah dengan melakukan analisis sentimen. Namun, dalam melakukan analisis sentimen sering didapati masalah dalam pemrosesan data teks yaitu gaya penulisan masyarakat di media sosial. Karena adanya batasan karakter dalam pembaharuan teks serta bebas menulis menggunakan bahasa apapun, maka didapati

tweet tidak menggunakan struktur kata yang baik dan benar atau juga penggabungan beberapa bahasa dalam satu kalimat.

Hasil penelitian yang dilakukan oleh Mohd Suhairi Md Suhaimin dan penelitian dari Mona Cindo yang membandingkan beberapa metode klasifikasi pada sentimen analisis menunjukan bahwa metode *Lexicon-based* memberikan hasil yang terbaik dari metode-metode lainnya [3][4]. Sehingga pada penelitian ini, peneliti menggunakan *text mining* pengklasifikasian analisis sentimen berupa teks dengan mengambil data dari sosial media *Twitter* terkait salah satu situs pembajak artikel penelitian yaitu, Sci-Hub, yang hasilnya adalah informasi berupa sentimen positif, negatif dan netral.

II. TINJAUAN PUSTAKA

A. Analisis Sentimen

Analisis sentimen atau *opinion mining* adalah bidang studi yang menganalisis pendapat, sentimen, evaluasi, sikap, dan emosi orang-orang terhadap suatu objek seperti produk, layanan, organisasi, individu, isu, peristiwa, topik dan lainnya [5]. Analisis sentimen yang juga disebut dengan *opinión mining* berfokus utama pada opini yang mengungkapkan atau menyiratkan sentimen positif atau negatif, tapi juga opini yang bersifat netral [6]. Pengaplikasian sentimen análisis telah menyebar ke hampir setiap bidang yang memungkinkan, mulai dari produk untuk konsumen, layanan, perawatan kesehatan, dan layanan keuangan hingga acara sosial dan politik. Dengan kegunaan serta pengaruh dari sentimen análisis di berbagai bidang yang bersangkutan menyebabkan banyak penelitian mengenai sentimen análisis sekarang ini. Kegunaannya menjadi acuan untuk mengembangkan suatu produk dan layanan yang baik dalam dunia usaha bisnis [5].

B. Twitter

Twitter adalah layanan pengiriman pesan yang memiliki banyak kesamaan dengan alat komunikasi yang sekarang ini digunakan. *Twitter* memiliki elemen yang mirip dengan *email*, IM, SMS, *blogging*, RSS, dan sebagainya. Tetapi beberapa faktor, terutama kombinasi, membuat *Twitter* unik. Pesan yang dikirimkan dan diterima di *Twitter* tidak lebih dari 140 karakter, yang artinya *Twitter* mudah digunakan untuk menulis dan membaca. Pesan-pesan di *Twitter* bersifat publik, yang artinya bertemu dengan orang baru di *Twitter* sangatlah mungkin terjadi [7]. Walaupun sekarang ini ada fitur *private* yaitu dimana orang-orang yang ingin mengikuti kita harus mendapatkan izin dari kita untuk mereka bisa melihat konten-konten *Twitter* milik kita, tapi berinteraksi dengan orang baru dengan *platform* ini sangatlah mungkin. Serta juga pesan yang dikirimkan dan diterima di *Twitter* sekarang ditambah menjadi 280 karakter. Pada situsnya, *Twitter* didefinisikan sebagai apa yang terjadi di dunia dan apa yang sekarang sedang orang-orang bicarakan, apa yang terjadi itu juga yang terjadi di *Twitter*, melihat apa yang orang-orang sedang bicarakan.

C. Metode Lexicon-Based

Indikator sentimen yang paling penting adalah kata sentimen, kata yang mengandung opini. Misalnya, baik, luar biasa, dan hebat adalah kata-kata bersentimen positif, sedangkan buruk, jelek, dan rusak adalah kata-kata sentimen negatif. Selain kata-kata individu, ada juga frase dan idiom, misalnya lengan dan kaki. Kata dan frase sentimen sangat penting untuk analisis sentimen. Daftar kata dan frasa semacam itu disebut sentimen *lexicon* [6]. Metode *lexicon-based* menggunakan *dictionary* atau kamus yang telah memuat nilai polaritas (positif, negatif) untuk penilaian pada kata-kata. *Dictionary* pada metode *lexicon-based* bisa dibuat secara manual [8].

D. Sci-Hub

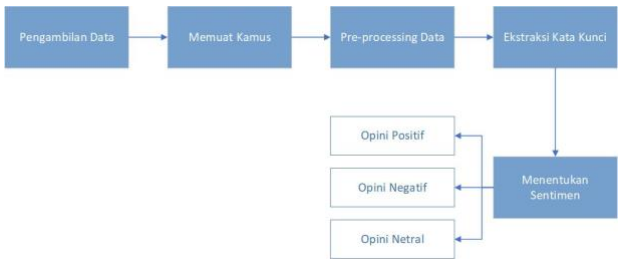
Sci-Hub merupakan situs *web* yang menyediakan akses ke literatur ilmiah melalui unduhan PDF teks lengkap. Situs memungkinkan pengguna untuk mengakses artikel yang seharusnya *paywalls* (berbayar). Sejak didirikan pada tahun 2011, Sci-Hub telah berkembang pesat dalam popularitas. Situs *web* Sci-Hub, sekarang dalam keberadaannya menyediakan akses gratis ke literatur ilmiah, meskipun kehadiran *paywalls* terus berlanjut. Sci-Hub menyebut dirinya sebagai "situs pembajak pertama di dunia untuk menyediakan akses massal dan publik ke puluhan juta makalah penelitian." Situs *web* dimulai pada tahun 2011, dijalankan oleh Alexandra Elbakyan, seorang mahasiswa pascasarjana dan penduduk asli Kazakhstan yang sekarang tinggal di Rusia [9].

E. Python

Python adalah bahasa pemrograman multguna. Tidak seperti bahasa lain yang susah untuk dibaca dan dipahami, Python lebih menekankan pada keterbacaan kode agar lebih mudah untuk memahami sintaks. Hal ini membuat Python sangat mudah dipelajari baik untuk pemula maupun untuk yang sudah menguasai bahasa pemrograman lain. Bahasa ini muncul pertama kali pada tahun 1991, dirancang oleh seorang bernama Guido van Rossum. Sampai saat ini Python masih dikembangkan oleh *Python Software Foundation.* Bahasa Python mendukung hampir semua sistem operasi, bahkan untuk sistem operasi Linux, hampir semua distronya sudah menyertakan Python di dalamnya [10].

III. METODE

Metode yang akan digunakan pada penelitian ini merupakan metode *Lexicon-based* yang memanfaatkan kamus dalam menilai polaritas suatu kalimat. Alur metode secara umum didefinisikan ada 5 tahapan [11]. Gbr. 1 merupakan tahapan penelitian dengan menggunakan metode *lexicon-based* secara umum:



Gbr. 1 Tahapan metode *lexicon-based* secara umum

A. *Teknik Pengumpulan Data*

Data diambil dari *Twitter* menggunakan *library* Python Twint dengan kata kunci "sci hub". Data yang diambil adalah data yang berbahasa Indonesia dan berbahasa Inggris. Data hasil pengambilan data atau *crawling* kemudian dibuat menjadi daftar pendapat atau opini yang setelahnya data akan diproses pada tahap *pre-processing*.

B. *Memuat Kamus*

Kamus merupakan komponen penting dalam menggunakan metode *Lexicon-based* [11]. Tahap ini memuat kamus yang berisi kata-kata yang sudah memiliki nilai sentimen yang akan digunakan pada saat tahap ekstraksi kata kunci dan penentuan sentimen. Kamus yang digunakan adalah kamus kata kunci positif dan kamus kata kunci negatif berbahasa Indonesia.

C. *Pre-processing*

Proses ini bertujuan untuk mempersiapkan data sebelum diproses di tahapan ekstraksi kata kunci dan menentukan sentimen. Berikut proses-proses yang dilakukan pada tahap ini:

- *Punctuation Removal* atau proses yang memisahkan tanda baca dan simbol lainnya selain alfabet, yang tujuannya agar tanda baca dan simbol-simbol tersebut tidak terproses saat proses dan tahap selanjutnya.
- Menghapus atribut *URL*, *Hashtag* dan *Mention* @user dari data *tweet*
- *Translation*, menerjemahkan data yang berbahasa Inggris ke bahasa Indonesia
- *Case Folding*/*Lowercase*, proses yang menjadikan setiap huruf besar ke huruf kecil agar tidak terjadi kesalahan pencocokan karakter atau huruf dalam suatu kata.
- Normalisasi Kata, mengubah kata-kata yang disingkat dan kata-kata gaul menjadi kata-kata dengan penulisan yang sesuai.
- *Stopwords Removal*, proses ini menghilangkan kata-kata yang tidak penting atau kata-kata yang tidak mengandung opini.
- Tokenisasi, memecahkan kalimat sehingga menjadi token-token yang akan dijadikan *keyword* atau kata kunci.

D. *Ekstraksi Kata Kunci*

Proses ini bertujuan untuk mengekstraksi kata-kata kunci untuk menentukan nilai dari kata tersebut positif atau negatif. Tahap ini merupakan tahap yang menggunakan kamus (*lexicon-based*) untuk mengidentifikasi kata ada atau masuk dalam kategori bernilai positif atau bernilai negatif. Data hasil *pre-process* dicocokkan dengan kamus yang akan digunakan untuk menentukan nilai polaritas setiap kata.

E. *Menentukan Sentimen*

Proses ini bagian dari menentukan sentimen kalimat, penentu dari sentimen dilakukan dengan menjumlahkan nilai-nilai kata kunci positif dan kata kunci negatif sesuai dengan penilaian dalam kamus. Proses ini juga merupakan tahapan akhir yang hasil dari pada tahap ini sudah merupakan hasil akhir. Berikut formula dalam menentukan sentimen:

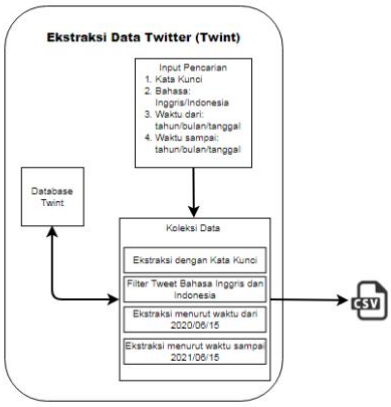
$$nilai\ sentimen = \begin{cases} 1, \sum sentimen > 0 \\ 0, \sum sentimen = 0 \\ -1, \sum sentimen < 0 \end{cases} \quad (1)$$

IV. HASIL DAN PEMBAHASAN

Hasil dari penelitian ini merupakan implementasi dari pada metode *lexicon-based* dalam menganalisis sentimen masyarakat pengguna sosial media *Twitter* terhadap situs pembajak artikel penelitian yaitu 'Sci-Hub'. Berikut tahapan yang telah dilakukan:

A. *Pengambilan Data*

Data diambil dari sosial media *Twitter* dengan menggunakan *library* dari Python yaitu *Twitter Intellegence Tools* (Twint) dengan alur proses pada Gbr. 2. Twint merupakan *tool* untuk pengambilan data yang ditulis dengan bahasa pemrograman Python yang memungkinkan untuk mengambil data dari *Twitter* tanpa menggunakan *Twitter* API [12]. Data yang diambil langsung dari *database* Twint sesuai dengan *input* pencarian yang hasil pengambilan data berformat *Comma Separated Values* (csv).



Gbr. 2 Alur Proses Pengambilan Data Twint.

Pengumpulan data dilakukan dengan cara *coding* menggunakan bahasa pemrograman Python di aplikasi Anaconda Jupyter Notebook. Gbr. 3 merupakan *source code* untuk pengambilan data menggunakan *library* Twint.

```
import twint

c = twint.Config()
c.Search = "scihub"
c.Since = 2020-06-15
c.Until = 2021-06-15
c.Store_csv = True
c.Output = "data.csv"
c.Lang = "en" #dan "in"

twint.run.Search(c)
```

Gbr. 3 Kode Pengambilan Data

Kata kunci yang dipilih adalah 'sci-hub' dengan kode bahasa Indonesia (in) dan bahasa Inggris (en) pada rentang waktu 15 Juni 2020 sampai 15 Juni 2021. Jumlah data yang terkumpul adalah 37,745 data dengan data kode bahasa Inggris sebanyak 28,194 dan kode bahasa Indonesia sebanyak 7,551 yang disimpan dalam format csv.

B. Menentukan Sentimen

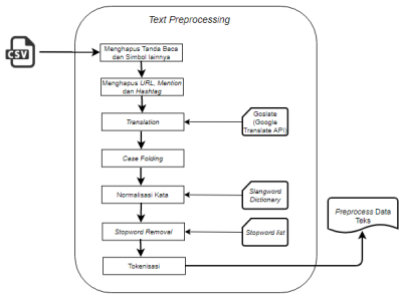
Kamus yang digunakan merupakan kamus yang disusun oleh Fajri K dan Gemala Y. R yang disebut dengan InSet [13]. Kata-kata dalam kamus sudah terlabel dan terbobot dalam label positif dan negative serta dibobot dengan polaritas angka positif (misalnya, 1) dan angka negatif (misalnya, -1). Seluruh kata-kata dalam kamus berjumlah 10,218. Tabel 1 merupakan beberapa kata dalam kamus:

TABEL I. ISI KAMUS INSET

Kata	**Bobot (*Weight*)**	**Jumlah Kata (*Number of Words*)**
gratis	4	1
makasih	5	1
sungguh-sungguh	2	2
tidak masuk akal	-4	3
sengsara	-5	1

C. Pre-processing

Tahap ini melakukan persiapan data yang telah diambil dari sosial media *Twitter* sebelum masuk ke tahap penentuan sentimen. Tahap ini memiliki beberapa proses yang dilakukan. Gambaran alur proses pada tahap ini adalah sebagai berikut:



Gbr. 4 Alur *Pre-processing*

Mengikuti alur proses pada Gbr 4, berikut merupakan proses-proses yang dilakukan:

1) Penghapusan Tanda Baca dan Simbol lainnya.

Tanda baca atau *punctuation* serta simbol lainnya selain alfabet pada semua data *tweet* dihapus, karena hanya data teks yang akan diproses untuk mengetahui sentimen. Proses ini dilakukan dengan menggunakan perintah pada Gbr. 5.

```
tweet = re.sub('[^a-zA-Z]', ' ', tweet)
tweet = re.sub("(\\d|\\W)+"," ",tweet)
tweet = re.sub(r'â', '', tweet)
tweet = re.sub(r'€', '', tweet)
tweet = re.sub(r'¦', '', tweet)
```

Gbr. 5 Kode Penghapusan Tanda Baca dan Simbol lainnya

Hasil dari proses ini dapat dilihat pada Tabel 2.

TABEL 2. HASIL PENGHAPUSAN TANDA BACA DAN SIMBOL LAINNYA

Sebelum	Sesudah
I Believe information should be free. So of course I side with Sci-Hub	I believe information should be free So of course I side with Sci hub
@schole_id Sudah banyak orang yg termudahkan risetnya, apalagi mahasiswa. Terimakasih Sci-Hub ðŸ˜	@schole_id Sudah banyak orang yg termudahkan risetnya apalagi mahasiswa Terima kasih Sci Hub

2) Penghapusan *Url*, *Mention* dan *Hashtag*

Data *tweet* yang diambil pada saat *crawling* memiliki atribut yang tidak perlu diproses pada saat menentukan sentimen, seperti *url*, *mention* dan *hashtag*. Sehingga atribut-atribut tersebut perlu untuk dihapus dengan tujuan agar tidak terproses pada proses selanjutnya. *Code* untuk penghapusan *url*, *mention*, dan *hashtag* ada pada Gbr. 6.

```
tweet = re.sub(r'https*\S+', ' ', tweet)
tweet = re.sub(r'@\S+', ' ', tweet)
tweet = re.sub(r'#\S+', ' ', tweet)
```

Gbr. 6 Kode Penghapusan *Url*, *Mention* dan *Hashtag*

Tabel 3. merupakan hasil dari penghapusan *url*, *mention* dan *hashtag*:

TABEL 3. HASIL PENGHAPUSAN *URL*, *MENTION* DAN *HASHTAG*

Sebelum	Sesudah
I Believe information should be free So of course I side with Sci Hub	I believe information should be free So of course I side with Sci hub
@schole_id Sudah banyak orang yg termudahkan	Sudah banyak orang yg termudahkan risetnya

risetnya apalagi mahasiswa Terimakasih Sci Hub	apalagi mahasiswa Terima kasih Sci Hub

3)*Translation*

Karena kamus InSet yang digunakan berisi kata-kata yang berbahasa Indonesia, data yang diterjemahkan dipisahkan kode *language* 'en' dan 'in'. Sehingga hanya data *tweet* dengan kode *language* 'en' yang berbahasa Inggris diterjemahkan ke bahasa Indonesia dengan menggunakan *library* Python Goslate. Goslate merupakan *library* Python untuk layanan terjemahan gratis yang mengakses Google *Translate* API dengan *querying* situs Google *Translate* [14]. Proses dilakukan dengan kode perintah pada Gbr. 7.

```
import goslate

with open('data.csv') as f:
    novel_text = f.read()

gs = goslate.Goslate()
translations = gs.translate(novel_text, 'id')
```

Gbr. 7 Kode Terjemahan Data

Adapun tujuan pada proses ini agar seluruh data *tweet* sudah berbahasa Indonesia untuk nantinya dapat dicocokan dengan kamus pada tahapan selanjutnya. Namun pada tahap ini, kata-kata yang masih berbahasa Inggris pada data dengan kode language 'in' dilewatkan. Tabel 4 merupakan beberapa contoh hasil terjemahan *tweet* dengan menggunakan *library* Python Goslate:

TABEL 4. HASIL TERJEMAHAN

Sebelum	Sesudah
I Believe information should be free So of course I side with Sci Hub	Saya percaya informasi harus gratis Jadi tentu saja saya berpihak pada Sci hub
Never use It is unfair to the profiteers	Jangan pernah menggunakan Ini tidak adil bagi pencatut

4)*Case Folding*

Proses ini mengubah setiap huruf menjadi *lowercase* atau huruf kecil dengan kode perintah pada Gbr. 8. Tujuannya agar setiap karakter teks sama sehingga tidak terjadi kesalahan dalam pencocokan teks.

```
tweet = tweet.lower()
```

Gbr. 8 Kode *Case Folding*

Beberapa contoh hasil perubahan pada proses ini ada pada Tabel 5.

TABEL 5. HASIL *CASE FOLDING*

Sebelum	Sesudah
Saya percaya informasi harus gratis Jadi tentu saja saya berpihak pada Sci hub	saya percaya informasi harus gratis jadi tentu saja saya berpihak pada sci hub
Sudah banyak orang yg termudahkan risetnya apalagi mahasiswa Terima kasih Sci Hub	sudah banyak orang yg termudahkan risetnya apalagi mahasiswa terima kasih sci hub

5)Normalisasi Kata

Pada proses ini data *tweet* dari hasil *crawling* ditemukan kata-kata yang disingkat-singkat dan tidak sesuai dengan penulisan kata yang baik sehingga perlu untuk kata-kata tersebut diubah menjadi kata-kata yang sesuai. *Dictionary* normalisasi kata diambil dari NLP Bahasa Indonesia *Resources* [15]. Kode perintah yang digunakan seperti pada Gbr. 9.

```
def clean_tweets(tweet):
    file_2  = open("slang_words.txt", "r")
    content2 = file_2.read()
    slang_words = ast.literal_eval(content2)
    file_2.close()


    word_tokens = word_tokenize(tweet)
    for w in word_tokens:
        if w in slang_words.keys():
            word_tokens[word_tokens.index(w)] = slang_words[w]
```

Gbr. 9 Kode Normalisasi Kata

Penggunaan *slangword dictionary* ini hanya untuk kata-kata yang berbahasa Indonesia sehingga kata-kata yang disingkat dan kata-kata gaul yang masih didapati berbahasa Inggris karena tidak terubah ketika proses terjemahan dilewatkan. Proses ini juga masih didapati ada beberapa kata yang berbahasa Indonesia tidak terubah sehingga penulis menambahkan secara manual ke dalam *dictionary*. Tabel 4 adalah hasil perubahan dari proses normalisasi kata:

TABEL 4. HASIL *CASE FOLDING*

Sebelum	Sesudah
sudah banyak orang yg termudahkan risetnya apalagi mahasiswa terima kasih sci hub	sudah banyak orang yang termudahkan risetnya apalagi mahasiswa terima kasih sci hub
cari jurnal gratis selain scihub ada di mana lagi ya ini scihub gue error mulu dari kemaren	cari jurnal gratis selain scihub ada di mana lagi ya ini scihub aku error terus dari kemarin

6)*Stopword Removal*

Pada proses ini setiap kata yang tidak mengandung opini atau tidak memiliki makna akan dihapus. Sehingga pada saat penentuan sentimen data yang diproses adalah kata-kata yang memang mengandung suatu opini. *List stopwords*

diambil dari NLP Bahasa Indonesia *Resources* [15]. Hasil perubahan saat penghapusan *stopword* ada pada Tabel 6.

TABEL 6. HASIL *STOPWORD REMOVAL*

Sebelum	Sesudah
sudah banyak orang yang termudahkan risetnya apalagi mahasiswa terima kasih sci hub	termudahkan risetnya mahasiswa terima kasih sci hub
cari jurnal gratis selain scihub ada di mana lagi ya ini scihub aku error terus dari kemarin	cari jurnal gratis scihub scihub error

7)Tokenisasi

Proses ini mengubah data *tweet* yang berupa kalimat menjadi beberapa bagian atau dalam bentuk token-token. Acuan sebagai antar pemisah token biasanya adalah spasi atau tanda baca. Hasil dari proses ini ada pada Tabel 7.

TABEL 7. HASIL TOKENISASI

Sebelum	Sesudah
termudahkan risetnya mahasiswa terima kasih sci hub	termudahkan ─ risetnya ─ mahasiswa ─ terima ─ kasih ─ sci ─ hub
cari jurnal gratis scihub scihub error	cari ─ jurnal ─ gratis ─ scihub ─ scihub ─ error

D. Ekstraksi Kata Kunci

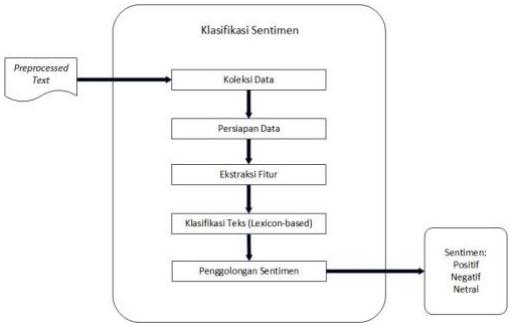
Setelah data melewati tahapan *pre-process*, data hasil *pre-process* yang berupa token-token akan diproses dengan mencocokkan token-token dengan kata-kata yang ada dalam kamus kata kunci positif dan kata kunci negatif. Pada penelitian ini, kamus InSet yang telah dimuat pada tahap sebelumnya kemudian dicocokkan dengan data hasil *pre-process*. Hasil dari pencocokkan tersebut belum merupakan hasil akhir atau nilai akhir sentimen.

TABEL 8. HASIL EKSTRAKSI KATA KUNCI

Data *Pre-process*	Kata Kunci Positif	Kata Kunci Negatif
termudahkan risetnya mahasiswa terima kasih sci hub	termudahkan, terima, kasih	
cari jurnal gratis scihub scihub error	gratis	error

E. Menentukan Sentimen

Data hasil dari pencocokkan dengan kamus kemudian dihitung nilai sentimennya. Jika hasil perhitungan data dominan nilai sentimen positif, maka sentimen kalimat tersebut positif. Sedangkan jika hasil perhitungan data dominan nilai sentimen negatif, maka sentimen kalimat tersebut adalah negatif. Namun jika hasil perhitungan nilainya sama, maka nilai sentimen kalimat tersebut adalah netral.



Gbr. 10 Alur Klasifikasi Sentiemen

Mengikuti Alur pada Gbr 10, data *pre-process* dipanggil dan dipersiapkan kembali atau kembali dinormalisasikan seperti pada tahap *pre-processing*. Tahapan ini juga menghapus data-data duplikasi atau data yang sama dan juga menghapus kata kunci 'sci-hub' sehingga tidak terproses pada saat pembuatan *wordcloud* kata sering muncul. Sehingga data berkurang menjadi 31,083 data.

Proses selanjutnya ekstraksi fitur, proses ini telah dijelaskan pada tahap sebelumnya dan contoh hasil bisa dilihat pada Tabel 8.

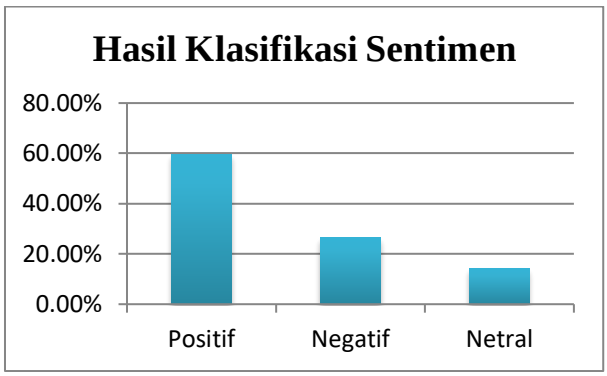
Setelah ekstraksi fitur, proses klasifikasi dimulai dengan menghitung data hasil pencocokkan kata kunci dengan nilai sentimen yang ada dalam kamus. Jika kata ada dalam kamus, kata tersebut diberi nilai yang sesuai dengan kamus. Proses tersebut dilakukan pada setiap kata dalam kalimat dan kemudian dihitung nilainya, jika jumlah nilai lebih besar dari angka nol (0) maka kalimat dikategorikan sentiment positif, sebaliknya jika jumlah lebih kecil dari angka nol (0) maka kalimat dikategorikan sentimen negatif, dan jika jumlah nilai sama dengan nol (0) maka kalimat dikategorikan sentimen netral.

TABEL 9. HASIL *DETERMINE SENTIMENT*

Tweet	∑KKP	∑KKN	Sen Pos	Sen Neg	Sen Net
termudahkan risetnya mahasiswa terima kasih sci hub	8		✓		
cari jurnal gratis scihub scihub error	4	-5		✓	

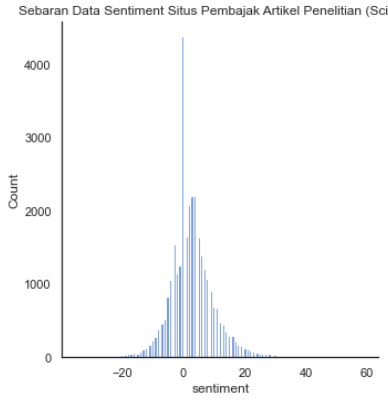
TABEL 10. HASIL KLASIFIKASI SENTIMEN DATA *TWITTER*

Sentimen	Jumlah (Angka)	Jumlah (%)
Positif	18,469	59.51%
Negatif	8,213	26.42%
Netral	4,374	14.07%
Total *Tweet*	31,083	100%



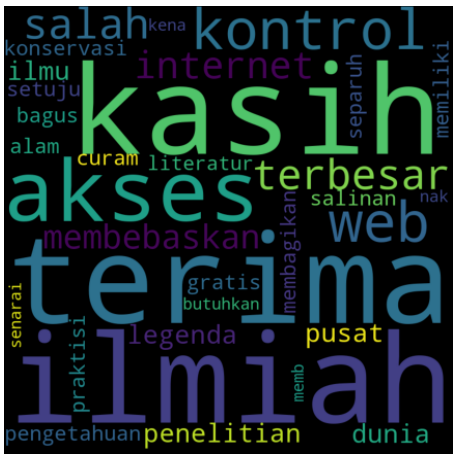
Gbr. 11 Hasil Klasifikasi Sentimen

Tabel 10 merupakan hasil dari klasifikasi sentimen data *Twitter* menggunakan metode *lexicon-based* dan Gbr 11 merupakan visualisasi grafik dari hasil klasifikasi sentimen. Bisa dilihat dari hasil klasifikasi, sentimen positif lebih besar dari sentimen negatif dan sentimen netral, yang artinya opini masyarakat mengenai situs pembajak artikel penelitian memberi respon positif. Gbr. 12 adalah grafik sebaran data nilai sentimen:



Gbr. 12 Sebaran Data Nilai Sentimen

Terdapat kata-kata yang sering muncul yang dominan bersentimen positif, seperti ‘terima’, ‘kasih’, ‘akses’, ‘gratis’. Kata-kata tersebut merupakan ekspresi pengguna sosial media *Twitter* menanggapi adanya situs pembajak artikel penelitian Sci-Hub. Maka dari itu dibuat visualisasi *wordcloud* kata sering muncul pada Gbr. 13 sehingga bisa melihat kata-kata apa saja yang sering muncul tentang situs pembajak artikel penelitian Sci-Hub.



Gbr. 13 *Wordcloud* Kata Sering Muncul

F. Pengujian dan Validasi

Dalam penelitian ini, proses validasi menggunakan perhitungan kesalahan klasifikasi secara manual. Penentuan kesalahan dilakukan dengan membandingkan nilai kelas yang sebenarnya dan hasil label dari klasifikasi sistem. Jika kelas klasifikasi sistem sama dengan label kelas sebenarnya, maka hasil klasifikasi dinyatakan benar. Akan tetapi, sebaliknya, jika kelas klasifikasi sistem tidak sama dengan label kelas sebenarnya, maka hasil klasifikasi dinyatakan salah.

Data pada tahap pengujian ini menggunakan data uji sebanyak 3108 data, yaitu 10% dari data keseluruhan hasil *pre-processing* dengan setiap data yang diambil mewakili *tweet* dengan waktu setiap hari pada data satu tahun. Data kelas sebenarnya dilabel secara manual oleh penulis dengan kelas label positif, negatif, dan netral.

Persamaan yang digunakan dalam tahap ini adalah sebagai berikut:

$$Akurasi = \frac{\sum pb}{num_data} \times 100\% \qquad (2)$$

Dimana:

pb : hasil klasifikasi sistem yang sesuai dengan label sebenarnya

num_data : jumlah seluruh data klasifikasi

Di bawah ini merupakan hasil perhitungan akurasi dengan menggunakan persamaan pada (2):

$$Akurasi = \frac{1730}{3108} \times 100\%$$

$$Akurasi = 0.556 \times 100\%$$

$$Akurasi = 55.6\%$$

Hasil perhitungan akurasi yang didapatkan adalah sebesar 55.6% untuk analisis sentimen menggunakan metode *lexicon-based* dengan 1730 data diprediksi benar oleh *classifier* dari 3108 data.

V. KESIMPULAN

Dari penelitian ini dapat disimpulkan hasil analisis sentimen situs pembajak artikel peneltiain dalam hal ini adalah Sci-Hub pada rentang waktu 15 Juni 2020 sampai 15 Juni 2021 dengan menggunakan metode *lexicon*-based memiliki sentimen positif sebesar 59,51%, sentimen negatif sebesar 26,42% dan sentimen netral 14,07%. Hal ini menunjukan bahwa sentimen positif lebih besar daripada sentimen negatif dan netral. Hasil pengujian akurasi sebesar 55,6% untuk 3108 data, dimana ada 1730 data yang berhasil diprediksi benar oleh sistem. Penelitian ini diharapkan dapat bermanfaat dan menjadi referensi bagi peneliti yang meneliti di bidang *data mining*, khususnya *text mining*. Kedepannya penelitian ini dapat dikembangkan dengan melakukan proses tokenisasi lainnya seperti bigram dan trigram, juga melabeli seluruh data untuk pengujian metode.

Kemudian pada penelitian ini untuk proses normalisasi kata, bisa lebih memperhatikan bilamana ada kata-kata yang tidak ada dalam *dictionary*. Kata bisa ditambahkan secara manual ke dalam *dictionary*. Proses ini juga bisa menambahkan *dictionary* normalisasi kata berbahasa Inggris agar setiap kata-kata bisa terproses. Pada tahapan *pre-processing* bisa menambahkan untuk mendeteksi majas untuk hasil yang lebih bagus.

Selain itu, penelitian kedepannya dapat menambahkan atau mengambil data dengan bahasa lainnya selain data berbahasa Indonesia dan bahasa Inggris untuk diolah.

DAFTAR PUSTAKA

[1] D. J. Fiander, "*Social Media for Academics*," *Chandos Publishing*, 2012.

[2] J. E. Sembodo, E. B. Setiawan dan ZK A. Baizal, "*Data Crawling* Otomatis pada *Twitter*," *Ind. Symposium on Computing*, 2016.

[3] M. S. Md Suhaimin, M. H. A. Hijazi, R. Alfred dan F. Coenen, "*Natural Language Processing Based Features for Sarcasm Detection: An Investigation Using Bilingual Social Media Texts*," *ICIT 2017 - 8th International Conference on Information Technology*, 2017.

[4] M. Cindo, D. P. Rini dan Ermatita, "*Literature Review*: Metode Klasifikasi Pada Sentimen Analisis," *SAINTEKS*, 2019.

[5] B. Liu, "*The Problem of Sentiment Analysis*," *Morgan & Claypool*, 2012.

[6] B. Liu, "*Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*," *Cambridge University Press*, 2015.

[7] T. O'Reilly and S. Milstein, "*The Twitter Book*," *O'Reilly Media*, 2011.

[8] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede, "*Lexicon-Based Methods for Sentiment Analysis*," *Association for Computational Linguistics*, 2011.

[9] D. S. Himmelstein, A. R. Romero, J. G. Levernier, T. A. Munro, S. R. McLaughlin, B. G. Tzovaras and C. S. Greene, "*Sci-Hub provides access to nearly all scholarly literature*," *eLife*, 2018.

[10] S. Hokya, "Buku Panduan Pemrograman Python," *Buku*, vol.84, hal. 487–492, 2013.

[11] A. Nurfalah, Adiwijaya dan A.A Suryani, "Analisis Sentimen Berbahasa Indonesia dengan Pendekatan *Lexicon-Based* pada Media Sosial," *Jurnal Masyarakat Informatika Indonesia*, 2017.

[12] PyPi, "Twint 2.1.20," *http://pypi.org/project/twint/*, 2020.

[13] F. Koto dan G. Y. Rahmaningtyas, "*InSet lexicon*: *Evaluation of a word list for Indonesian sentiment analysis in microblogs*," *International Conference on Asian Language Processing*, 2017.

[14] PyPi, "Goslate 1.5.1," *https://pypi.org/project/goslate*, 2016.

[15] GitHub, "NLP Bahasa Indonesia *Resources*," *https://github.com/louisowen6/NLP_bahasa_resources*, 2020.

[16] R. Arief dan K. Imanuel, "Analisis Sentimen Topik Viral Desa Penari pada Media Sosial *Twitter* dengan Metode *Lexicon Based*," *MATRIK*, 2019.

[17] Statiswaty, Rusnia dan N. Ransi, "Analisis Sentimen Wisata Bahari di Sulawesi Tenggara Memanfaatkan Media Sosial *Twitter* dengan Menggunakan Metode *Lexicon-Based*," *semanTIK*, 2017.