

Journal Artificial Intelligence, Multimedia, and Mobile Technology (AI2MTech)

Homepage jurnal: <https://journal.darmajaya.ac.id/index.php/AI2MTech>

Penerapan *Lexicon-Based* Untuk Analisis Sentiment Masyarakat Terhadap Kasus Kebocoran Data Di Indonesia

Rahmalia Syahputri¹, Juan Khrisna Tanubrata², Sherli Trisnawati³

Fakultas Ilmu Komputer, Informatics & Business Institute Darmajaya

Jl. Z.A Pagar Alam No 93, Bandar Lampung – Indonesia 35142 Telp (0721) 787214 fax. (0721) 700261

corresponding author: rahmalia@darmajaya.ac.id

ARTICLE INFO

Article history:

Submitted 14 August 2024

Received 15 August 2024

Received in revised form 09 September

2024 Accepted 15 November 2024

Available online on 27 Desember 2024

Keywords:

Lexicon-Based Approach, Cybercrime,
Data Leakage, Dictionary-Based
Approach, Sentiment Analysis.

Kata kunci:

Lexicon-based approach, Kejahatan siber,
Kebocoran Data, Dictionary-Based
Approach, Analisa sentimen.

ABSTRACT

The growth of social media has increased the availability of user-generated content that can be utilized for sentiment analysis, providing valuable insights into how the public perceives major social issues. One of the most critical issues widely discussed is cybercrime, particularly in the form of data leaks, which threaten individual privacy, institutional trust, and national security. Data leaks are increasingly recognized as part of a broader cybercrime landscape that demands public awareness as well as effective responses from stakeholders. Understanding how society reacts to this issue is therefore essential for government institutions and businesses. This study applies a lexicon-based method using three different sentiment dictionaries to classify sentiments expressed on Twitter regarding data leak issues. Data acquisition was conducted through web scraping, followed by preprocessing stages including case-folding, tokenization, stopword removal, and stemming. Sentiment classification was then performed to categorize tweets into positive, negative, and neutral classes. The results indicate that positive sentiment dominates with 2,866 tweets (45.07%), followed by negative sentiment with 2,298 tweets (36.12%), and neutral sentiment with 1,191 tweets (18.73%). These findings demonstrate that lexicon-based sentiment analysis, supported by multiple dictionaries, provides more reliable and comprehensive results than a single lexicon.

ABSTRAK

Pertumbuhan media sosial telah meningkatkan ketersediaan konten yang dihasilkan pengguna yang dapat dimanfaatkan untuk analisis sentimen, sehingga memberikan wawasan berharga mengenai persepsi publik terhadap isu-isu besar. Salah satu isu krusial yang banyak dibicarakan adalah kejahatan siber, khususnya dalam bentuk kebocoran data, yang mengancam privasi individu, kepercayaan institusi, serta keamanan nasional. Kebocoran data semakin diakui sebagai bagian dari lanskap kejahatan siber yang lebih luas, sehingga membutuhkan kesadaran publik sekaligus respons yang efektif dari berbagai pemangku kepentingan. Memahami reaksi masyarakat terhadap isu ini menjadi sangat penting baik bagi lembaga pemerintah maupun dunia usaha. Penelitian ini menerapkan metode berbasis leksikon dengan menggunakan tiga kamus sentimen yang berbeda untuk mengklasifikasikan sentimen yang diekspresikan di Twitter terkait isu kebocoran data. Akuisisi data dilakukan melalui web scraping, diikuti tahap praproses seperti case folding, tokenisasi, penghapusan stopword, dan stemming. Klasifikasi sentimen kemudian dilakukan untuk mengelompokkan tweet ke dalam kategori positif, negatif, dan netral. Hasil penelitian menunjukkan bahwa sentimen positif mendominasi dengan 2.866 tweet (45,07%), diikuti sentimen negatif dengan 2.298 tweet (36,12%), dan sentimen netral dengan 1.191 tweet (18,73%). Temuan ini membuktikan bahwa analisis sentimen berbasis leksikon, yang diperkuat dengan penggunaan beberapa kamus, dapat memberikan hasil yang lebih andal dan komprehensif dibanding hanya menggunakan satu kamus.

1. Pendahuluan

Cybercrime adalah industri triliunan yang mengintai pengguna internet. Di Indonesia, pada awal tahun 2019, beberapa percobaan serangan dilakukan terhadap Bukalapak yang merupakan marketplace terbesar ketiga [1]. Selanjutnya, pada Mei 2020, Tokopedia, yang merupakan marketplace terbesar kedua, dilaporkan 91 juta data pengguna telah terjual di dark web [2]. Tidak hanya pada tahun sebelumnya, pada November tahun 2021 dilaporkan dua database yang berisikan data personel polri bocor dan diekspos oleh peretas dengan mengunggahnya ke platform social media twitter[3]. Di tahun 2022 juga terjadi kebocoran data sebanyak 400.000 data SIM card diretas [4].

Ditengah kondisi terjadinya kebocoran data saat ini tingkat kepercayaan masyarakat kepada pemerintah menjadi berkurang secara signifikan. Berdasarkan data dari CNN Indonesia, sebanyak 46% masyarakat ragu dengan data pribadi mereka, sebanyak 43% masyarakat masih percaya dengan keamanan data pribadi mereka, 10% dari masyarakat kurang percaya, dan 1% masyarakat tidak percaya sama sekali data pribadi mereka akan dikelola dengan baik oleh penyedia layanan digital [5]. Memahami bagaimana opini masyarakat dapat membantu institusi untuk mengambil keputusan[6]. Hal ini memungkinkan untuk melakukan analisa dan penelitian lebih lanjut mengenai presentase tingkat kepercayaan masyarakat terhadap Pemerintah dan pelaku usaha berbasis digital berdasarkan komentar atau postingan di media sosial.

Text mining adalah langkah dalam analisis teks yang dilakukan secara otomatis oleh komputer untuk mengekstrak informasi berkualitas tinggi dari sekumpulan teks yang disusun menjadi sebuah dokumen[7]. Text mining juga mencakup pengetahuan tentang ekstraksi sebuah teks dan juga pengolahan data teks untuk berbagai aplikasi dan juga penelitian [8][9][10] Untuk melakukan perhitungan mengenai presentase tingkat masyarakat, penelitian ini menggunakan sebuah teknik yaitu sentiment analysis. Sentiment analysis adalah salah satu teknik yang umum digunakan untuk mengungkapkan fenomena sosial berbasis komputasi, teknik ini juga merupakan teknik komputasi text mining berbasis natural language processing (NLP)[11]. Sentimen sendiri sendiri merupakan istilah yang digunakan untuk menggambarkan topik subjektif dan objektif, faktual atau non-faktual, dengan opini positif atau negatif[12]. Ada dua pendekatan yang umum digunakan dalam teknik analisa sentimen yaitu pendekatan berbasis machine learning dan pendekatan berbasis lexicon [13][14]. Penggunaan machine learning menawarkan keunggulan yaitu peran setiap kata dalam proses kategorisasi sentimen disesuaikan dengan korpus dan aplikasi. Namun saat korpus yang tersedia tidak cukup banyak, seperti pada kasus penelitian ini, maka diperlukan pendekatan berbasis lexicon untuk kategorisasi sentimen [13].

Fokus pada penelitian ini adalah untuk menganalisa sentiment masyarakat dengan menggunakan pendekatan lexicon. Dataset yang diambil dari media sosial Twitter yang cukup trend di kalangan Masyarakat untuk memberikan opini baik berupa tanggapan maupun kritik [15][16]. Tanggapan atau kritik cukup beragam dari opini atas suatu mesin kendaraan [6] hingga kandidat politik sebelum melakukan pencoblosan suara [17]. Hal ini mengeser kebiasaan di masyarakat yang sebelumnya adalah bertanya dengan keluarga atau teman [6][17].

Penelitian tentang sentiment analysis dilakukan oleh [18]. Penelitian ini membuat sebuah kamus InSet Lexicon yang berisikan Indonesia sentiment lexicon dengan tujuan untuk mengidentifikasi dalam penulisan opini dan mengkategorikan termasuk opini positif atau negatif. Penelitian ini menghasilkan 3.609 kata positif dan 6.609 kata negatif. Setiap kata pada penelitian ini dilabelkan secara manual berdasarkan polaritasnya kemudian ditingkatkan dengan menambahkan stemming dan synonym set. Hasil akurasi dari penelitian ini adalah 65,78%.

Pada penelitian[19], penelitiannya menerapkan tiga langkah dalam preprocessing yaitu menghilangkan stopword, mengubah bahasa gaul menjadi kata baku berdasarkan Kamus Besar Bahasa Indonesia (KBBI), dan menghilangkan kata jenis subjek atau objek. Hasil dari penggunaan tiga langkah tersebut meningkatkan akurasi yang cukup signifikan sebesar 3% hingga 3,5%. Jika menggunakan fitur

unigram dan ekstraksi fitur Count-Vectorizer maka akan mencapai akurasi sebesar 88%. Berdasarkan penelitian[20], evaluasi pada penelitian ini dilakukan pada 2 tingkat, tingkat dokumen dan tingkat kalimat. Penelitian ini juga menggunakan pendekatan lexicon dengan InSet lexicon. Hal ini dilakukan dengan tujuan untuk menguji seberapa baik dampak InSet lexicon terhadap analisis sentimen Bahasa Indonesia. Hasil penelitian ini mencapai akurasi sebesar 90,9%, namun hasil yang lebih baik diperoleh pada tataran kalimat setelah melakukan perubahan pada kamus lexicon.

Penelitian [21] menghadirkan sebuah fitur baru yang dinamakan lexico-semantic. Dalam penelitiannya lexico-semantic memiliki fungsi untuk memberikan bantuan dan meningkatkan dalam proses klasifikasi sentimen. Hasil yang di dapatkan dalam penelitian ini adalah PP(Precision-positive) 0.788, PN(Precision-negative) 0.822, FI-Positive 0.745, FI-Negative 0.864, RP(Recall-Positive) 0.720, dan RN (Recall-Negative) 0.891. hasil tersebut juga dibandingkan dengan hasil dari penelitian [22] dengan hasil PP(Precision-positive) 0.714, PN(Precision-negative) 0.829, FI-Positive 0.663, FI-Negative 0.856, RP(Recall-Positive) 0.621, dan RN (Recall-Negative) 0.885. Hal ini membuktikan, penggunaan fitur lexico-semantic memberikan peningkatan yang cukup baik dan memiliki hasil yang lebih konsisten.

Dalam penelitiannya, [23] melakukan pelabelan data menggunakan vader-lexicon, preprocessing menggunakan metode text mining dan pembobotan term frekuensi menggunakan TF-ID. Hasil pembobotan menggunakan TF-ID adalah 49% (489 ulasan) sentimen positif, 37% (366 ulasan) sentiment netral, dan 14% (145 ulasan) sentimen negatif. Dari hasil perbandingan 1000 sample data menggunakan vader lexicon terkait dengan data ulasan PLN mobile menghasilkan 67% positif, 6% netral dan 27% negatif. Pada penelitian ini juga membandingkan dengan menggunakan naïve bayes dengan perbandingan data latih 90:10 menggunakan confusion matrix, mendapatkan hasil akurasi sebesar 70%.

Penelitian dengan menggunakan algoritma lexicon-based juga dilakukan oleh [24]. Dalam penelitiannya, dilakukan analisis terkait trending topics. Sebanyak 45.198 data tweet didapatkan dan diolah dalam penelitiannya. Nilai akurasi pengujian 80% menunjukan nilai yang baik dan nilai rata-rata yang dihasilkan dari keseluruhan topik adalah 100% akurasi. Penelitian tentang analisis sentiment juga dilakukan oleh [25]. Dalam penelitian ini, menyajikan penggunaan Formal Concept Analysis (FCA) untuk membuat sebuah kamus untuk analisis sentimen berbasis dokumen text. Hasil akurasi yang di dapatkan dalam penelitian ini adalah 0.820.

Penelitian lainya tentang analisis sentimen dilakukan oleh [26]. Penelitian ini menggunakan automatic lexicon Senti-N-Gram. Penggunaan leksikon Senti-N-Gram mempengaruhi hasil analisis sentimen kebijakan new normal dengan menggunakan objek bahasa Indonesia. hasil accuracy dalam penelitian dihasilkan 63,42%, precision sebesar 77%, recall sebesar 62,88% dan f-measure sebesar 69,32%. Adapun hasil akurasi terbaik dengan menggunakan lexicon Senti-N-Gram sebesar 0,6342. Hasil terbaik tersebut menggunakan data sebanyak 350 data latih, dan data uji, yang terbagi dari 229 tweet positif dan 121 tweet negatif.

Berkaitan dengan penelitian yang telah dilakukan sebelumnya, maka dilakukan penelitian dengan menggunakan topik yang lebih baru karena topik ini merupakan hal yang masih belum banyak diteliti. Penelitian ini akan menganalisa sentimen masyarakat terhadap kasus kebocoran data di Indonesia dengan menggunakan metode lexicon-based approach. Salah satu perbedaan pada penelitian ini adalah perluasan kamus/dictionary yang digunakan. Penelitian ini menambahkan kamus/dictionary yang dinamai dengan kamus serangan yang berisikan kata-kata yang terkait dengan aktivitas kebocoran data.

Penulisan artikel ini dibagi menjadi empat bagian. Pada bagian pertama, kami memaparkan latar belakang dan penelitian terkait yang menjadi landasan penelitian ini. Pada metode penelitian yang merupakan bagian kedua dari artikel ini, kami menyampaikan tahapan penelitian yang kami lakukan. Sedangkan hasil penelitian, kami sampaikan di bagian hasil dan pembahasan, dan bagian terakhir adalah Kesimpulan.

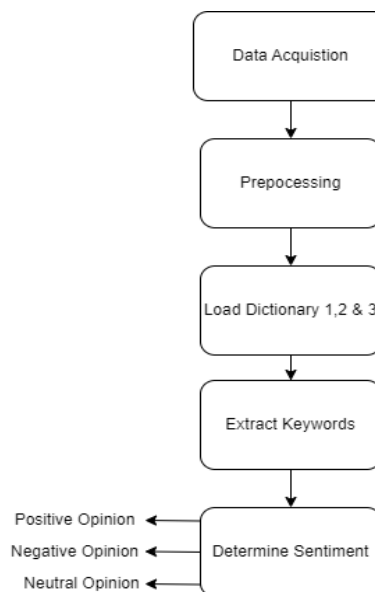
2. Metode Penelitian

Penelitian ini akan mengelompokkan tanggapan atau komentar masyarakat yang ada di media sosial yang akan di analisa dan diubah menjadi data sentiment. Objek pada penelitian ini adalah mengklasifikasikan data sentiment tersebut apakah termasuk kategori tanggapan positif, negatif atau netral berdasarkan batasan atau aturan yang diberikan. Data tanggapan atau komentar masyarakat pada media sosial yang akan diklasifikasikan adalah data yang diposting di media sosial twitter periode Agustus – September 2022.

Dalam tahapan penelitian ini terbagi menjadi 5 tahapan, yaitu: Data acquisition, Preprocessing, Load Dictionary 1, 2 & 3, extract keywords, dan determine sentiment sebagaimana digambarkan pada gambar 1. Penelitian ini melibatkan tiga kamus yang dibangun dan menambah jumlah kamus yang ketiga untuk memasukkan kata khusus yang terkait dengan aktivitas serangan pada jaringan komputer seperti denial of service (DoS) dan hack. Alasan lain modifikasi adalah untuk meminimalkan kekeliruan dalam menentukan opini. Hal ini disebabkan adanya tantangan dalam penelitian analisa sentiment, antara lain kata-kata yang digunakan untuk mengungkapkan sentimen bersifat khusus untuk domain tertentu [27], misalnya kata “berhasil” dan “membangun” adalah kata positif, namun jika digabungkan dengan kata “Hack” atau “DoS” maka pembentukan kalimat “Berhasil Hack” akan menghasilkan opini negatif.

Data Acquisition

Pada tahapan ini dilakukan pengambilan data yang diambil dari data tweets pada media sosial <https://twitter.com> dengan menggunakan query pencarian mengenai kasus kebocoran data yang terjadi di Indonesia. Tweets yang di ambil mengandung hastag #bjorka, #Bjorka, #BlokirKominformo, #Hacker, #Jebol, #Bocor, #Tersebar, #Tersiar, #Terbongkar, #Terekspose, #Terungkap, #Terkuak, #Tembus. Data acquisition diambil dengan menggunakan pengkodean yang dibangun dengan menggunakan Bahasa pemrograman Python.



Gambar 1. Tahapan Penelitian

Preprocessing

Pada tahapan preprocessing ini dilakukan cleaning data dalam dataset yang di ambil pada langkah sebelumnya. Didalam dataset yang telah di ambil memiliki beberapa attribute kata yang perlu dihapus seperti: tanda baca, angka, hastag (#), dan ad (@). Tahapan penghapusan tersebut termasuk kedalam

tahapan cleaning dalam preprocessing. Tahapan cleaning terbagi menjadi beberapa tahapan yaitu: penghapusan karakter, pergantian kata, stemming, dan tokenization.

1) Penghapusan Karakter

Pada tahapan preprocessing penghapusan karakter ini, dilakukan penghapusan karakter seperti: tanda baca, angka, hastag (#), dan add (@). Contohnya: Ente kadang kadang lemah banget sistemnya @DivHumas_Polri yang benerin nanti @kemkominfo tambah lemah #opposite6890 #BrigadirYosua #brigaderjoshua. diubah menjadi Ente kadang kadang lemah banget sistem benerin lemah opposite BrigadirYosua Brigaderjoshua.

2) Pergantian Kata

Pada tahapan ini, dilakukan pergantian kata dilakukan untuk mengubah kata kata yang mengandung huruf kapital menjadi huruf kecil. Contohnya: Ente kadang kadang lemah banget sistemnya benerin lemah opposite BrigadirYosua Brigaderjoshua, diubah menjadi ente kadang kadang lemah banget sistem benerin lemah opposite brigadirYosua brigaderjoshua.

3) Stemming

Pada tahapan stemming ini, semua kata-kata yang sebelumnya sudah diubah menjadi huruf kecil diubah kembali menjadi kata dasar. Tahapan ini dilakukan dengan menggunakan library pada Python yaitu library Sastrawi. Library ini dipergunakan karena memiliki fungsi untuk mengubah kalimat berbasis bahasa Indonesia yang di yang digunakan dalam penelitian kedalam kata dasar. Contoh: Kominform diluar nalar sangat tidak asik pilih kasih blokir nya blokirkominform diubah menjadi kominform, luar, nalar, asik, pilih, kasih, blokir, blokirkominform.

4) Tokenization

Tahapan tokenization ini merupakan tahapan terakhir dalam proses preprocessing. Pada tahapan ini dilakukan pemecahan kalimat menjadi kata. Hal ini dilakukan untuk memudahkan dalam proses perbandingan pada kamus yang digunakan dan menentukan berapa nilai dari masing-masing kata tersebut berdasarkan kamus tersebut. Contohnya: main rtp maxwin tinggi rilis kokotogel diproses menjadi [main, rtp, maxwin, tinggi, rilis, kokotogel].

Load Dictionary

Pada proses load dictionary dilakukan pemuatan tiga kamus yaitu kamus positif, negatif dan juga serangan. Masing-masing kamus memiliki bobot nilai tersendiri. untuk jangkauan nilai pada kamus bernilai dari -5 hingga +5. Nilai pada kamus berfungsi untuk menentukan berapa jumlah score yang didapatkan berdasarkan kata-kata yang dianalisis.

Extract Keywords

Dalam tahapan ini, dilakukan ekstrasi kata berdasarkan hasil preprocessing yang berupa kata yang sudah di-tokenization. Contohnya: main rtp maxwin tinggi rilis kokotogel diproses menjadi [main, rtp, maxwin, tinggi, rilis, kokotogel].

Determine Sentiment

Tahapan ini merupakan tahapan penentuan hasil klasifikasi sentiment berdasarkan score yang didapatkan dari hasil perbandingan masing-masing kata dengan kamus yang digunakan. Pada penelitian ini hasil klasifikasi dibagi menjadi 3 klasifikasi yaitu: sentimen positif, sentiment negatif, dan sentiment netral.

3. Hasil dan Pembahasan

3.1. Data Acquisition

Data tweet yang telah dikumpulkan tentang kebocoran data adalah berupa dataset yang berisikan 15.000 data. Semua data tersebut diambil dari tweet yang mengandung hastag #bjorka, #Bjorka, #BlokirKominfo, #Hacker, #Jebol, #Bocor, #Tersebar, #Tersiar, #Terbongkar, #Terekspos, #Terungkap, #Terkuak, #Tembus. Contoh data dapat dilihat pada tabel 1.

Tabel 1. Contoh Data

NO	USER	TWEET
1	Andrew_felix93	Kominfo diluar nalar,sangat tidak asik,pilih kasih blokir nya #blokirkominfo 🤔🤔 https://t.co/X8Cguao63h Ente kadang kadang lemah banget sistemnya @DivHumas_Polri yang benerin nanti @kemkominfo tambah lemah #opposite6890
2	ganosecteam01	#BrigadirYosua #brigaderjoshua @Opposite6980 #brigaderj #UstadzKamiBukanTeroris #boikotkominfo #BoikotKominfo #BlokirKominfo #blokirkomifo #AparatBobrok

Hasil dari tabel 1 didapatkan dari hastag yang terkandung didalamnya. Contohnya: untuk akun 1, cuitan tersebut mengandung hastag #BlokirKominfo. Sehingga data cuitan tersebut termasuk dalam dataset yang digunakan dalam penelitian ini.

3.2. Preprocessing

Pada penelitian ini *preprocessing* dilakukan menggunakan empat tahapan. Masing – masing tahapan pada *preprocessing* ini bertujuan untuk *cleaning* dataset yang sudah di ambil sebelumnya. Misalnya untuk postingan “Ente kadang kadang lemah banget sistemnya @DivHumas_Polri yang benerin nanti @kemkominfo tambah lemah #opposite6890 #BrigadirYosua #brigaderjoshua @Opposite6980 #brigaderj #UstadzKamiBukanTeroris #boikotkominfo #BoikotKominfo #BlokirKominfo #blokirkomifo #AparatBobrok <https://t.co/kkqrxVH8X>”, maka tahapannya ada pada tabel 2.

Tabel 2. Tahapan Preprocessing

Penghapusan Kata	Ente kadang kadang lemah banget sistemnya yang benerin nanti tambah lemah opposite BrigadirYosua brigaderjoshua Opposite brigaderj UstadzKamiBukanTeroris boikotkominfo BoikotKominfo BlokirKominfo blokirkomifo AparatBobrok
Pergantian Kata	ente kadang kadang lemah banget sistem benerin lemah opposite brigadiyosua brigaderjoshua opposite brigaderj ustadzkamibukanteroris boikotkominfo boikotkominfo blokirkominfo blokirkomifo aparatbobrok
Stemming	ente, kadang, lemah, banget, sistem, benerin, lemah, opposite, brigadiyosua, brigaderjoshua, opposite, brigaderj, ustadzkamibukanteroris, boikotkominfo ,boikotkominfo, blokirkominfo, blokirkomifo, aparatbobrok
Tokenization	[ente, kadang, lemah, banget, sistem, benerin, lemah, opposite, brigadiyosua, brigaderjoshua, opposite, brigaderj, ustadzkamibukanteroris, boikotkominfo ,boikotkominfo, blokirkominfo, blokirkomifo, aparatbobrok]

Dapat dilihat pada tabel 2, text asli berisikan kalimat. Pada tahap pertama dilakukan proses penghapusan kata. Proses penghapusan kata ini bertujuan untuk menghapus beberapa atribut yang terkandung dalam kalimat seperti; @, # dan juga link yang terkandung dalam kalimat. Tahap kedua adalah pergantian kata. Dalam tahapan ini berfungsi mengubah seluruh kata yang mengandung huruf kapital menjadi huruf kecil. Tahapan ketiga dilakukan proses *Stemming*, tahapan ini berfungsi untuk menghapus beberapa imbuhan-imbuhan pada kalimat. Contohnya kalimat kadang-kadang diubah menjadi kadang. Untuk tahapan akhir, dilakukan proses *tokenization*. Proses ini bertujuan untuk membagi kalimat utuh menjadi token. Contoh “Saya mau makan” yang setelah di tokenasi menjadi “saya”, “mau”, “makan”.

3.3. Load Dictionary

Load dictionary tiga kamus yaitu kamus positif (Tabel 3) dan negatif (Tabel 4), dan kamus serangan (Tabel 5). kata yang telah diolah pada tahap preprocessing akan di bandingkan dengan masing-masing kamus. Tiap kamus memiliki range nilai yang berbeda-beda yang bernilai -5 sampai +5. Khusus untuk kamus serangan bobot yang diberikan adalah -5, hal ini dikarenakan kata yang ada dalam kamus serang merupakan kata yang dipergunakan dalam menyebut istilah serangan dalam jaringan komputer dan kejahatan siber. Sehingga sangat berkaitan dengan topik pada penelitian ini dan tersebut tergolong kata negatif.

Tabel 3. Contoh Data Kamus Positif

Kata	Bobot Nilai
Hai	3
Merekam	2
Ekstensif	3
Paripurna	1

Tabel 4. Contoh Data Kamus Negatif

Kata	Bobot Nilai
Rampok	-5
Anjir	-3
Satir	-2
Bodo	-5

Tabel 5. Contoh Data Kamus Serangan

Kata	Bobot Nilai
Serang	-5
Dos	-5
Ddos	-5
Hack	-5

Dalam Tabel 3, 4, dan 5 merupakan contoh dari masing-masing kamus yang digunakan dalam penelitian ini. Tabel 3 dan Tabel 4 merupakan kamus yang berasal dari penelitian sebelumnya dari [18]. Kemudian Tabel 5 adalah kamus baru yang dibuat dan berisikan kata-kata yang berkaitan dengan aktivitas penyerangan yang merupakan bagian dari serangan siber.

3.4 Extract Keywords

Dalam tahapan ini dilakukan pengeluaran kata yang sudah didapatkan sebelumnya (Tabel 6). Extract keywords dilakukan dengan tujuan untuk mengeluarkan hasil preprocessing pada tahapan tokenisasi. Kemudian akan dilanjutkan pada tahap *determine sentiment*.

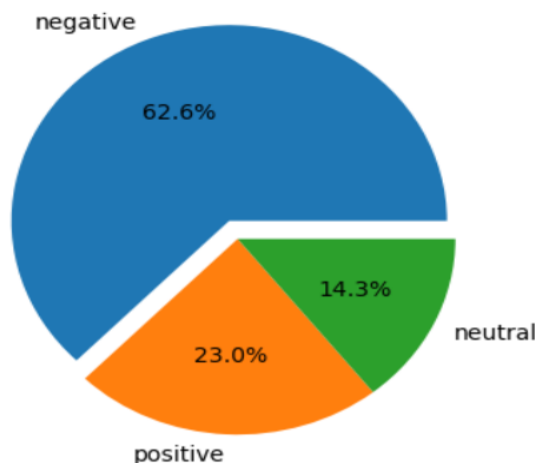
Tabel 1. Contoh Extract Keywords

Tweets	Tweets Token
ente, kadang, lemah, banget, sistem, benerin, lemah, opposite, brigadirjosua, brigaderjoshua, opposite, brigaderj, ustadzkamibukanteroris, boikotkominfo ,boikotkominfo, blokirkominfo, blokirkomifo, aparatbobrok	[ente, kadang, lemah, banget, sistem, benerin, lemah, opposite, brigadirjosua, brigaderjoshua, opposite, brigaderj,ustadzkamibukanteroris, boikotkominfo ,boikotkominfo, blokirkominfo, blokirkomifo, aparatbobrok]

3.5 Determine Sentiment

Tahapan ini dilakukan untuk menentukan klasifikasi sentiment. Penentuan hasil klasifikasi sentiment didapatkan berdasarkan jumlah skor yang di dapatkan dari tiap-tiap kata yang dibandingkan berdasarkan kamus. Gambar 2 memperlihatkan bahwa mayoritas masyarakat mengekspresikan sentimen negatif terhadap kasus kebocoran data. Hal ini menunjukkan rendahnya tingkat kepercayaan publik terhadap kemampuan pemerintah dalam melindungi data pribadi. Sentimen positif relatif kecil (23%), menandakan bahwa dukungan atau apresiasi terhadap kinerja pemerintah hanya muncul dalam isu-isu tertentu. Sementara itu, 14,3% netral mengindikasikan adanya kelompok masyarakat yang bersikap hati-hati atau tidak cukup informasi untuk menilai.

Sentiment Polarity on Tweets Data



Gambar 1 Hasil Determine Sentiment

Tabel 7. Contoh Sentimen Positif

Tweets	Polarity_S core	Polarity
main rtp maxwin tinggi rilis kokotogel blokirkominfo agustus paypal steam kokotogel promo bonuscashback slotonline togelonline slotgaco	7	Positif

ente kadang kadang lemah banget sistem benerin lemah opposite brigadiyosua brigaderjoshua opposite brigaderj ustadzkamibukanteroris boikotkominfo boikotkominfo blokirkominfo blokirkomifo aparatbobrok	4	Positif
besok event hadiah hp samsung galaxy blokirkominfo agustus paypal steam kokotogel promo bonuscashback slotonline togelonline slotgacor	4	Positif
syarat photocard bts tokopedia viral gempa worldwar ikatancinta news islam muslim haram kpop oppa vaksin indonesia peoplepower korea kdrama hacker tauhid artis	13	Positif
event hadiah fantastik tunggu infona besok blokirkominfo agustus paypal steam kokotogel promo bonuscashback slotonline togelonline slotgacor	3	Positif

Pada Tabel 7, sebagian besar sentimen positif berasal dari tweet yang berhubungan dengan promosi, event, atau hadiah, yang secara tidak langsung dikaitkan dengan tagar #BlokirKominfo. Meskipun kontennya bukan kritik terhadap pemerintah, kata-kata yang bernuansa positif tetap terdeteksi. Hal ini menunjukkan bahwa konteks penggunaan hashtag tidak selalu relevan dengan isu kebocoran data, namun berpengaruh pada klasifikasi sentiment.

Tabel 2. Contoh Sentimen Negatif

Tweets	Polarity_S core	Polarity
maluuu bgttt ekspos	-4	Negatif
kominfo luar nalar asik pilih kasih blokir blokirkominfo	-5	Negatif
tanggap hacker putih 101aehyung101nal kait bocor foto pribadi jennie blackpink bts vbts kimtaehyung 101aehyung jennie kimjennie btsv btsarmy bts blackpink hacker rumor https co tlhbfxsjpb	-6	Negatif
mudah musim persib bobol lawan hayuu gaess baik luismilla tiket bobol banjir jebol bek fullbek strategi	-18	Negatif
kuak identitas laku aliando syarief alami emotional abuse domestic abuse paksa rodi ancam siksa sinergianews https co rognik cuaca cuacabmkg bmkg	-16	Negatif

Di tabel 8, Tweet dengan sentimen negatif umumnya berupa kritik langsung terhadap pemerintah atau institusi terkait, misalnya penggunaan kata ‘tolol’, ‘tidak becus’, hingga sindiran keras kepada Kominfo. Ada juga tweet yang mengaitkan isu kebocoran data dengan kasus publik lain (artis, olahraga, atau politik). Nilai polarity_score yang rendah (hingga -18) memperlihatkan akumulasi kata-kata negatif yang sangat kuat.

Tabel 3. Contoh Sentiment Netral

Tweets	Polarity_S core	Polarity
ygy blokirkominfo	0	Netral
ngak cepat kawan pln shaddy orang humanrights blokirkominfo	0	Netral
ngomong doank blokirkominfo blokirkapakemikir humanrights	0	Netral
kerja ni bubar bubar blokirkominfo	0	Netral
blokirkominfo woi buka gofud timeout	0	Netral

Tabel 9 menunjukkan bahwa Tweet netral biasanya bersifat informatif atau sekadar menyebut hashtag tanpa ekspresi opini. Contoh: 'ygy blokirkominfo' atau 'kerja ni bubar bubar blokirkominfo'. Netralitas muncul karena tidak ada kata dalam tweet yang terklasifikasi ke dalam kamus positif maupun negatif. Namun, hal ini juga mengindikasikan keterbatasan pendekatan lexicon, karena beberapa kalimat sindiran halus atau ironi tidak terbaca sebagai negatif.

Tabel 4. Hasil Analisa Ahli Bahasa.

Tweets	Hasil Analisa Ahli Bahasa
PERMAINAN BARU DENGAN RTP DAN MAXWIN TERTINGGI, SEGERA RILIS DI KOKOTOGEL!	Negatif
#BlokirKominfo Agustus PayPal Steam #kokotogel #promo #bonuscashback #slotonline #togelonline #slotgacor https://t.co/6UyqO49BCN Good morning, Indonesia ...	Negatif
Melindungi data sendiri aja kagak becus, belaga lagi mau terapkan PSE ** SaveDigitalFreedom ** PSEmelanggarHAM #HumanRights #BlokirKominfo #bersamakominfo #indonesiahancur https://t.co/DiQioFRIDa Terima kasih, 23.840 pendukung sudah teken petisi ini!	Positif
Namun belum ada kata terlambat, mari kita bersama-sama untuk Teken, Sebarkan & Viralkan!	
#BlokirKominfo #ProtesNetizen #BubarkanKominfo #BatalkanDNSNasional #PSEMelanggarHAM #SaveDigitalFreedom https://t.co/nQOWftRBr0 Maju terus! Tetap semangat! ðŸ’ªðŸ’»	Positif
#BlokirKominfo #BlokirGakPakeMikir #HumanRights #HumanRightsViolations #HumanRights #HumanRightsViolation #BlokirGakPakeMikir #BlokirKominfo #PSEMelanggarHAM #BatalkanDNSNasional #protesnetizen @hrw @EFF	Positif
Kominfo ini permainan sengaja bocor bocorkan data agar mereka mau memperketat PSE terus hujat! Emang tolol,kebiasaan buruk di jdiin pembelaan, klo emng gak tau tata krama ya "minta maaf" jgn nyari pembelaan klo lu paling bener, dan ngerti tata krama. #BlokirKominfo #Tolol	Negatif
Setuju Bang BPJ Desak Aparat Kepolisian Juga Kominfo Usut Tuntas Kebocoran Data Pelanggan PLN	Positif
#BlokirKominfo #DataPLNBocor	

Butuh yang mantap - mantap? ada di KOKOTOGEL! Dengan minimal deposit 10.000 anda bisa menikmati semua fasilitas permainan kami! Mulai dari SLOT, CASINO, TOGEL, DLL !	Negatif
#BlokirKominfo Agustus PayPal Steam #kokotogel #promo #bonuscashback #slotonline #togelonline #slotgacor https://t.co/QhpI0HMBZg @RaffiAditya998 @Kemlu_RI @MataNajwa @NarasiNewsroom @Menlu_RI @ukr_embassy @ZelenskyyUa @VHamianin Masalah ini bisa dilihat ke PSE Kominfo ahkir Juli lalu, dengan pemblokiran app vital untuk masyarakat seperti Steam dan Paypal, masyarakat membuat trend seperti #BlokirKominfo dan unjuk rasa, ujung-ujungnya membawa rasa tidak kepercayaan masyarakat. (Thread 3/4)	Positif
Haha kok Polda metro Jaya gini yah emang gak bisa amankan website nya gimana @kemkominfo kalau gaksalah inianggota Sambo bukan @Opposite6980 #opposite6890 #BrigadirJ #brigadirjoshua #BrigadirYosua #boikotkominfo #BoikotKominfo #BlokirKominfo #blokirkomifo #AparatBobrok https://t.co/QuszHo5WaN	Positif

Perbandingan pada Tabel 10 menunjukkan perbedaan yang signifikan antara hasil sistem dan penilaian ahli bahasa. Misalnya, tweet promosi seperti ‘PERMAINAN BARU ...’ terklasifikasi negatif oleh sistem karena mengandung kata terkait serangan (‘bocor,’ ‘hack’), namun ahli bahasa menilai konteksnya positif atau netral. Hal ini menjelaskan rendahnya akurasi (50%) dan menegaskan tantangan utama pendekatan lexicon, yakni kesulitan memahami konteks, sarkasme, dan ironi yang lazim di media sosial.

4. Kesimpulan

Penelitian ini menerapkan analisis sentimen berbasis lexicon untuk mengevaluasi opini publik terhadap kasus kebocoran data di Indonesia dengan menggunakan 15.000 tweet. Hasil analisis menunjukkan bahwa sentimen negatif mendominasi sebesar 62,6%, sedangkan sentimen positif sebesar 23,0% dan sentimen netral 14,3%. Penambahan kamus khusus serangan siber terbukti memperkaya klasifikasi sentimen, meskipun tingkat akurasi sistem dibandingkan dengan ahli bahasa masih berada pada angka 50%. Meski demikian, penelitian ini memiliki keterbatasan: data hanya diambil dalam periode singkat, pendekatan lexicon masih belum baik mendeteksi sarkasme atau makna kontekstual, serta tingkat akurasi relatif rendah. Temuan ini menegaskan rendahnya kepercayaan publik terhadap upaya pemerintah dalam mengamankan data pribadi serta pentingnya perumusan kebijakan keamanan siber yang lebih kuat. Penelitian selanjutnya dapat meningkatkan akurasi dengan mengintegrasikan pendekatan machine learning serta memperluas kamus domain khusus untuk menangkap nuansa linguistik seperti sarkasme dan makna kontekstual pada data media sosial.

REFERENCES

- [1] astrid rahadiani Putri and J. Yordan, “Bukalapak Akui 13 Juta Data yang Dijual Hacker adalah Peretasan di Maret 2019,” kumparan. Accessed: Sept. 01, 2023. [Online]. Available: <https://kumparan.com/kumparantech/bukalapak-akui-13-juta-data-yang-dijual-hacker-adalah-peretasan-di-maret-2019-1tMRTr1UR0G>
- [2] T. Citradi, “Data Pengguna E-Commerce Bocor, Ini Bukti RI Rawan Cybercrime,” CNBC Indonesia. Accessed: Oct. 21, 2022. [Online]. Available: <https://www.cnbcindonesia.com/tech/20200706140435-37-170521/data-pengguna-e-commerce-bocor-ini-bukti-ri-rawan-cybercrime>
- [3] L. J. Dicky Prastya, “Daftar Kasus Kebocoran Data di Indonesia selama 2021, Termasuk Sertifikat Vaksin Jokowi,” suara.com. Accessed: Aug. 10, 2023. [Online]. Available: <https://www.suara.com/tekno/2022/01/01/015822/daftar-kasus-kebocoran-data-di-indonesia-selama-2021-termasuk-sertifikat-vaksin-jokowi>
- [4] “3 tri.co.id,” BreachForums. Accessed: Oct. 30, 2022. [Online]. Available: <https://breached.to/>
- [5] CNN, “Masyarakat Indonesia Ragu Soal Pengelolaan Data Pribadi,” teknologi. Accessed: May 02, 2023. [Online]. Available: <https://www.cnnindonesia.com/teknologi/20190626180223-185-406756/masyarakat-indonesia-ragu-soal-pengelolaan-data-pribadi>
- [6] J. K. Rout, K.-K. R. Choo, A. K. Dash, S. Bakshi, S. K. Jena, and K. L. Williams, “A model for sentiment and emotion analysis of unstructured social media text,” *Electronic Commerce Research*, vol. 18, pp. 181–199, 2018.
- [7] Y. N. Prasetya and D. Winarso, “Penerapan Lexicon Based Untuk Analisis Sentimen Pada Twiter Terhadap Isu Covid-19,” vol. 11, p. 7, 2021.
- [8] C. Dreisbach, T. A. Koleck, P. E. Bourne, and S. Bakken, “A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data,” *Int J Med Inform*, vol. 125, pp. 37–46, May 2019, doi: 10.1016/j.ijmedinf.2019.02.008.
- [9] T. A. Koleck, C. Dreisbach, P. E. Bourne, and S. Bakken, “Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review,” *J Am Med Inform Assoc*, vol. 26, no. 4, pp. 364–379, Feb. 2019, doi: 10.1093/jamia/ocy173.
- [10] X. 1 Wang, J. 1 Tian, and F. 2 1 Z. H. I. T. C. Li, “Text data mining of power based on natural language processing technology,” p. 012050, May 2022, doi: 10.1088/1742-6596/2221/1/012050.
- [11] Rintyarna and B. Setya, “Joint Distribution pada Weighted Majority Vote (WMV) untuk Peningkatan Kinerja Sentiment Analysis Tersupervisi pada Dataset Twitter,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 9, no. 5, Art. no. 5, Oct. 2022.
- [12] M. Wongkar and A. Angdresey, “Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter,” Oct. 2019, pp. 1–5. doi: 10.1109/ICIC47613.2019.8985884.
- [13] C. S. Khoo and S. B. Johnkhan, “Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons,” *Journal of Information Science*, vol. 44, no. 4, pp. 491–511, Aug. 2018, doi: 10.1177/0165551517703514.
- [14] B. Setya Rintyarna, R. Sarno, and C. Fatichah, “Semantic Features for Optimizing Supervised Approach of Sentiment Analysis on Product Reviews,” *Computers*, vol. 8, no. 3, Art. no. 3, Sept. 2019, doi: 10.3390/computers8030055.
- [15] N. Agustina, A. Adrian, and M. Hermawati, “Implementasi Algoritma Naïve Bayes Classifier untuk Mendeteksi Berita Palsu pada Sosial Media,” *Faktor Exacta*, vol. 14, no. 4, Art. no. 4, Jan. 2022, doi: 10.30998/faktorexacta.v14i4.11259.
- [16] F. T. Saputra, Y. Nurhadryani, S. H. Wijaya, and D. Defina, “Analisis Sentimen Bahasa Indonesia pada Twitter Menggunakan Struktur Tree Berbasis Leksikon,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 8, no. 1, Art. no. 1, Feb. 2021, doi: 10.25126/jtiik.0814133.
- [17] V. Bonta, N. Kumares, and N. Janardhan, “A comprehensive study on lexicon based approaches for sentiment analysis,” *Asian Journal of Computer Science and Technology*, vol. 8, no. S2, pp. 1–6, 2019.

- [18] F. Koto and G. Rahmaningtyas, *InSet Lexicon: Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs*. 2017. doi: 10.1109/IALP.2017.8300625.
- [19] A. S. A. Siti Khomsah, "Text-Preprocessing Model Youtube Comments in Indonesian | Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)," Aug. 2020, Accessed: Aug. 25, 2023. [Online]. Available: <http://jurnal.iaii.or.id/index.php/RESTI/article/view/2035>
- [20] R. Firdaus, I. Asror, and A. Herdiani, "Lexicon-Based Sentiment Analysis of Indonesian Language Student Feedback Evaluation," *Indonesian Journal on Computing (Indo-JC)*, pp. 1-12 Pages, May 2021, doi: 10.34818/INDOJC.2021.6.1.408.
- [21] M. Mohd, S. Javeed, Nowsheena, M. A. Wani, and H. A. Khanday, "Sentiment analysis using lexico-semantic features," *Journal of Information Science*, p. 01655515221124016, Oct. 2022, doi: 10.1177/01655515221124016.
- [22] H. Saif, M. Fernandez, and H. Alani, "SentiCircles for Contextual and Conceptual Semantic Sentiment Analysis of Twitter," May 2014, pp. 83–98. doi: 10.1007/978-3-319-07443-6_7.
- [23] Y. Asri, W. N. Suliyanti, D. Kuswardani, and M. Fajri, "Analisis Sentimen Pelabelan Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam menganalisis sentimen data ulasan PLN Mobile: Analisis Sentimen," *PETIR*, vol. 15, no. 2, Art. no. 2, Dec. 2022, doi: 10.33322/petir.v15i2.1733.
- [24] Ismarmiaty, M. N. Asti, and A. A. Rizal, "Lexicon Based Sentiment Analysis pada Trending Topic di Nusa Tenggara Barat," *Jurnal Informatika dan Teknologi Komputer (J-ICOM)*, vol. 3, no. 2, Art. no. 2, Nov. 2022, doi: 10.33059/j-icom.v3i2.6136.
- [25] M. Ojeda-Hernández, D. López-Rodríguez, and Á. Mora, "Lexicon-based sentiment analysis in texts using Formal Concept Analysis," *International Journal of Approximate Reasoning*, vol. 155, pp. 104–112, Apr. 2023, doi: 10.1016/j.ijar.2023.02.001.
- [26] R. A. Siregar, Y. A. Sari, and I. Indriati, "Analisis Sentimen Kebijakan New Normal dengan Menggunakan Automated Lexicon Senti N-Gram," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 1, Art. no. 1, Feb. 2023, doi: 10.25126/jtiik.20231015006.
- [27] S. Agarwal and A. Sureka, "Applying Social Media Intelligence for Predicting and Identifying On-line Radicalization and Civil Unrest Oriented Threats," *arXiv:1511.06858 [cs]*, Nov. 2015, Accessed: Mar. 17, 2022. [Online]. Available: <http://arxiv.org/abs/1511.06858>