Klasterisasi Keyword Terkait Pornografi pada Media Sosial Twitter Menggunakan Latent Dirichlet Allocation

Qonita Nailul Muna[1], Rifda Awalia Zuhroh[2], Vania Rahma Dianutami[3], Nur Aini Rakhmawati[4]

[1, 2, 3, 4] Departemen Sistem Informasi, Institut Teknologi Sepuluh Nopember
[1]qonitamuna.19052@student.its.ac.id
[2]rifdazuhroh.19052@student.its.ac.id
[3]dianutami.19052@student.its.ac.id
[4]nur.aini@is.its.ac.id

Media sosial adalah saluran komunikasi bersifat *online* sebagai media yang digunakan untuk berbagi berbasis komunitas. Media sosial memungkinkan seseorang terhubung satu dengan yang lain tanpa harus bertemu secara tatap muka. Twitter merupakan salah satu media sosial yang memberikan kebebasan bagi para penggunanya untuk membuat, mengunggah, dan membaca unggahan yang disebut *tweet* dengan jumlah pengguna di Indonesia mencapai 18,45 juta di tahun 2022. Twitter ternyata ramai digunakan sebagai media penyebarluasan konten asusila seperti pornografi. Pada penelitian ini, penulis mencari tahu beberapa kata kunci yang sering digunakan dalam penyebar luasan konten pornografi di Twitter dengan menggunakan metode *Latent Dirichlet Allocation* (LDA) untuk menemukan topik yang dominan dari kata kunci yang digunakan dan mengelompokkan kata kunci secara otomatis. Pada penerapan LDA, fitur *stopword* digunakan untuk mengeliminasi kata-kata yang tidak diperlukan. Cara menentukan jumlah topik yang optimal yaitu dengan melihat nilai *perplexity* dan topik *coherence*. Dari total data 15.135 unggahan yang didapatkan dari *data crawling*, selanjutnya dipetakan menjadi lima topik dan topik yang paling banyak digunakan menggunakan kata kunci sange, nonton, dan bokepindo.

Kata Kunci— Latent Dirichlet Allocation*, Pornografi, Twitter, *Data Crawling*, *Stopword

I. PENDAHULUAN

Perkembangan teknologi informasi meningkat pesat dan berdampak pada setiap lapisan masyarakat. Semua orang dapat mengakses internet dan berseluncur di Media sosial dengan mudah. Media sosial adalah saluran komunikasi bersifat *online* sebagai media yang digunakan untuk berbagi berbasis komunitas. Media sosial merupakan hasil perkembangan teknologi dibidang komunikasi. Media sosial memungkinkan seseorang terhubung antara satu dengan yang lain tanpa harus bertemu secara tatap muka yang menyebabkan cepatnya penyebarluasan suatu informasi [1]. Dalam media sosial, konten positif bukan satu-satunya hal yang dapat ditemukan. Mereka juga akan menemukan konten negatif karena Media sosial seperti Twitter memberikan kebebasan kepada setiap orang untuk berekspresi secara bebas. Berdasar laporan We Are Social, pengguna Twitter di Indonesia mencapai 18,45 juta di tahun 2022 [2]. *Twitter* merupakan salah satu media sosial yang memberikan kebebasan bagi para penggunanya untuk membuat, mengirim, dan membaca pesan yang disebut *tweet*. Namun kenyataannya Twitter menjadi tempat yang ramai digunakan untuk penyebarluasan konten asusila seperti pornografi. Kementerian Komunikasi dan Informatika (Kominfo) menyebutkan bahwa banyak konten pornografi yang menyebar di Twitter yang mana sampai saat ini Kominfo aktif mencari memblokir situs yang berkaitan dengan konten pornografi [1][3]. Pada tahun 2015, Yuliandre Darwis mengungkapkan bahwa Indonesia berada di posisi kedua untuk kategori negara paling banyak mengakses konten pornografi dan sekitar 80% akun yang menyebarluaskan konten negatif di Twitter merupakan konten pornografi. Hal ini membuktikan bahwa pornografi merupakan konten negatif yang paling banyak tersebar di Media sosial khususnya Twitter [4].

Dalam Media sosial Twitter bertebaran bermacam-macam konten termasuk konten pornografi. Kominfo menyatakan bahwa konten pornografi paling banyak ditemukan di Media sosial Twitter. Jumlah akun yang menyebarkan konten pornografi ditemukan hingga mencapai 600.000 akun. Berdasarkan aduan konten negatif yang diterima oleh Kominfo, konten pornografi mendapatkan aduan terbanyak dengan total 244.738 konten sepanjang tahun 2019 [3] . Pola penyebaran konten pornografi melalui media sosial ini cenderung bebas dan pengguna media sosial lainnya merasa penyebaran konten pornografi ini membuat iklim pada media sosial tercemar akibat adanya penyebaran konten negatif tersebut. Bentuk dari konten pornografi bermacam-macam seperti yang disebutkan dalam Undang Undang nomor 44 tahun 2008, pornografi dalam gambar, sketsa, ilustrasi, foto, tulisan, suara, bunyi, gambar bergerak, animasi, kartun, percakapan, gerak tubuh, atau bentuk pesan lainnya melalui berbagai bentuk media komunikasi dan/atau pertunjukan di muka umum, yang memuat kecabulan atau eksploitasi seksual yang melanggar norma kesusilaan dalam masyarakat [5]. Perilaku pornografi dipicu beberapa faktor salah satunya adalah seberapa sering orang tersebut mengakses konten pornografi yang tampaknya digemari oleh kaum remaja. Adanya akses tayangan pornografi di media sosial yang saat ini marak terjadi di Indonesia mengakibatkan tingginya berbagai perilaku menyimpang yang melanggar nilai-nilai dan norma keasusilaan pada diri remaja yang berkembang secara terus menerus [6].

Menurut Ketua ikatan Sarjana Komunikasi Indonesia (ISKI), 80 persen pemuda Indonesia menyimpan konten pornografi di telepon genggamnya [3]. Dan tampaknya dalam hal ini, Media sosial telah menjadi sarana berbagi berbagai jenis konten pornografi yang dapat diakses siapa pun, kapan pun dan di mana pun [7]. Sementara itu, Indonesia merupakan negara hukum yang menjunjung tinggi nilai moral, etika, dan akhlak mulia serta hukum di Indonesia memiliki kedudukan yang kuat di mana dalam menyelesaikan suatu permasalahan mengedepankan hukum. Pornografi merupakan jenis konten negatif yang larangannya sudah diatur dalam undang-undang yang harus dipatuhi di Indonesia. Meskipun telah ada Undang Undang yang mengatur mengenai Pornografi, nyatanya masih banyak pihak yang menyebarluaskan konten pornografi di Media sosial. Sampai saat ini, masih banyak pelaku penyebarluasan konten pornografi di Twitter yang lolos dari pandangan hukum dan belum ditindaklanjuti untuk dilakukan pemblokiran atau bahkan penutupan akun [8]. Hal ini terbukti dari masih banyaknya konten terkait pornografi yang tersebar di Twitter dan akun yang menyebarluaskan konten tersebut masih aktif sampai saat ini.

Terdapat beberapa penelitian yang telah dilakukan sebelumnya berkaitan dengan penelitian yang dilakukan penulis. Salah satu penelitian yang dilakukan sebelumnya yaitu menggunakan bantuan *machine learning*, seperti *decision tree*, *naïve bayes*, dan *Support Vector Machines* (SVM). Penelitian yang pernah dilakukan sebelumnya ini bertujuan untuk membandingkan beberapa metode tersebut dan mencari tahu metode terbaik yang dapat digunakan untuk melakukan klasifikasi konten pornografi dengan Bahasa Indonesia dan Bahasa Inggris yang bertebaran di media sosial Twitter. Penulis penelitian tersebut juga melakukan percobaan tambahan untuk meningkatkan kinerja dalam aktivitas klasifikasi. Penelitian itu menunjukkan bahwa tingkat akurasinya cukup tinggi. Namun, penggunaan tata bahasa yang berbeda dapat menjadi suatu kendala yang memengaruhi keakuratan [9]. Selain itu, terdapat penelitian lainnya yang relevan dengan penelitian yang dilakukan oleh penulis yaitu melakukan identifikasi terhadap kalimat-kalimat pornografi yang muncul pada suatu artikel maupun *web page.* Penelitian ini menggunakan metode *machine learning* yang sama-sama bertujuan untuk memperoleh hasil perbandingan mengenai metode yang paling baik yang dapat digunakan untuk melakukan klasifikasi. Dalam penelitian ini, inti dari sistem yaitu mengklasifikasikan suatu kalimat ke salah satu dari dua kategori yaitu pornografi atau non-pornografi. Proses identifikasi dalam penelitian ini diuji menggunakan *K-Nearest Neighbor* (KNN), *passive aggressive classifier*, dan SVM dimana hasil dari penelitian ini yaitu algoritma SVM memiliki akurasi tertinggi yaitu sebesar 98,25% [10]. Selain dua penelitian tersebut, terdapat penelitian lainnya dimana penelitian ini menggunakan metodologi yang serupa dengan yang digunakan oleh penulis dalam melakukan penelitian ini yaitu *Latent Dirichlet Allocation* (LDA). Penelitian lainnya yang relevan yaitu penelitian yang hendak melakukan pengelompokan topik cuitan akun bot Twitter yang menggunakan tagar #Covid-19 di mana penelitian tersebut juga menggunakan LDA sebagai metode penelitian. Penulis penelitian tersebut menggunakan objek Twitter karena Twitter merupakan media sosial yang saat ini sedang pesat perkembangannya. Tagar *trending* pada media sosial Twitter dapat dengan cepat berubah karena cepatnya sebuah informasi tersebar luas dalam media sosial tersebut. Penelitian tersebut ingin mengetahui apa saja topik yang dibahas oleh akun bot mengenai Covid-19 mengingat informasi yang dapat tersebar tidak hanya berita baik, namun juga berita buruk, atau bahkan berita yang menggiring opini publik mengenai Covid-19. Penelitian tersebut berakhir menghasilkan lima topik teratas yang paling banyak dibahas di Twitter oleh akun bot yaitu antara lain kondisi dan dampak pandemi saat ini, himbauan untuk menjaga jarak agar kesehatan tetap terjaga, perkembangan penyebaran Covid-19 yang ada di Indonesia, vaksinasi yang terjadi di beberapa wilayah di Indonesia, dan cara menghadapi Covid-19 [11].

Berdasarkan permasalahan yang sudah dijelaskan sebelumnya, penelitian yang dilakukan ini bertujuan untuk melakukan analisis terhadap konten pornografi yang tersebar luas di salah satu Media sosial yaitu Twitter dan data yang diambil dari Media sosial berupa *tweet* yang mengandung konten pornografi. Penelitian ini dilakukan dengan salah satu metode *topic modelling* yang sedang populer saat ini yaitu metode *Latent Dirichlet Allocation* (LDA) yang dapat digunakan untuk mengetahui macam-macam kata kunci yang sering digunakan oleh pelaku penyebarluasan konten pornografi di Twitter dengan melakukan klasterisasi [12]. Dengan penelitian ini, diharapkan dapat membantu pihak manajemen KOMINFO yang mengharapkan Media sosial dapat dimanfaatkan untuk penggunaan hal-hal yang bersifat produktif.

II. TINJAUAN PUSTAKA

A. Topic Modelling

Topic modelling adalah sebuah metode *unsupervised machine learning* yang melakukan klasifikasi untuk menemukan variabel laten dari data teks besar. *Topic modelling* terdiri dari entitas-entitas berupa "kata" ataupun "dokumen". Ide dasar dari *topic modelling* adalah sebuah topik yang terdiri dari kata-kata tertentu yang menyusun topik tersebut, dan dalam satu dokumen memiliki kemungkinan terdiri dari beberapa topik-topik dengan probabilitas tertentu. Tujuan dari *topic modelling* adalah untuk menemukan topik dan kata yang tersembunyi dalam topik tersebut. Kumpulan dokumen biasanya terdapat beberapa distribusi topik, dengan distribusi topik tersebut dapat diketahui seberapa banyak penggambaran masing-masing topik tersebut terlibat dalam sebuah dokumen dan hal ini dapat diketahui topik utama mana yang dibahas dalam suatu dokumen [13]. Metode yang paling popular dalam *topic modelling* adalah *Latent Dirichlet Allocation* (LDA).

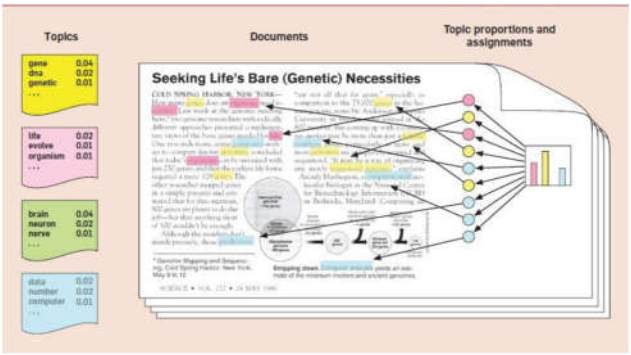
B. Text Mining

Text mining merupakan terminologi yang umum dijumpai pada berbagai studi literatur yang melibatkan data berbentuk tekstual dalam skala besar. Teknologi *text mining* meruapakan

bagian dari teknologi penambangan data (*data mining*). *Text mining* adalah suatu proses penambangan data berupa *text* dari data dan informasi yang bersumber dari dokumen. Tujuan dari *text mining* ini adalah untuk mengekstraksi kata-kata yang dapat mendeskripsikan sebagian besar isi dokumen sehingga data yang ditemukan dapat dianalisis hubungannya dengan dokumen lainnya [14]. Sumber data yang digunakan dalam *text mining* menggunakan sumber data yang berasal dari dokumen berbasis teks yang tidak memiliki struktur. Langkah pertama yang dilakukan dalam teknik *text mining* adalah *pre-processing*, lalu dilanjutkan proses *data mining*, dan diakhiri dengan proses *post-processing*. Penggunaan teknologi *text mining* sangat berguna dalam pengolahan data berjumlah besar karena proses analisis data tersebut dapat dilakukan serentak dalam kurun waktu tertentu.

C. Metode Latent Dirichlect Allocation

Latent Dirichlet Allocation (LDA) adalah sebuah *generative probabilistic model* untuk menemukan *latent semantic topic* dalam kumpulan *data text*. Metode ini pertama kali dipublikasikan pada tahun 1936 oleh Ronald A. Fisher melalui *paper The Use of Multiple Measure in Taxonomic Problems*. LDA dapat digunakan untuk mengelompokkan kata kunci secara otomatis. LDA biasanya digunakan untuk klasterisasi, melakukan peringkasan, menghubungkan, dan dapat memproses data dengan memberikan bobot pada masing-masing dokumen yang nantinya menghasilkan daftar topik. Masukan yang diterima oleh LDA dapat berupa judul-judul dokumen dan menghasilkan pola dari dokumen yang dimasukkan [15]. Oleh karena itu, semakin banyak dokumen yang dimasukkan, maka semakin akurat model yang dihasilkan pada algoritma.



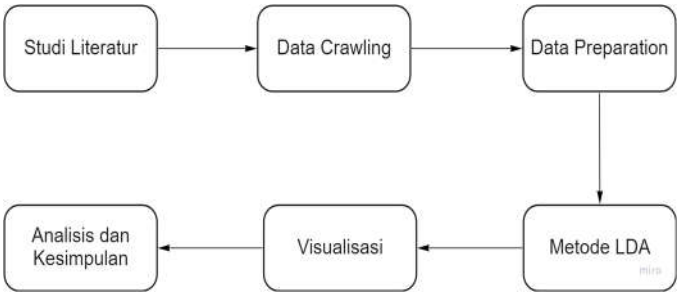
Gbr 1 Ilustrasi penggunaan LDA [15]

LDA menggunakan metode *bag of words* yang digunakan untuk mengidentifikasi topik tersembunyi dalam dokumen besar. Mekanisme kerja LDA dibagi menjadi dua, yaitu inferensi dan implementasi. Inferensi merupakan proses untuk menentukan pembobotan dari setiap kata yang terdapat di setiap dokumen dalam korpus. Implementasi merupakan tahapan lanjutan dari aplikasi LDA untuk memenuhi kebutuhan temu antar informasi selanjutnya. Model LDA dapat diimplementasikan menggunakan bahasa pemrograman Phyton dengan menggunakan *library genism*. Hasil luaran dari *package* tersebut berupa grafik yang sesuai dengan topik bahasan [16].

Pada penelitian sebelumnya sudah ada penelitian yang membahas mengenai klasterisasi menggunakan metode LDA. Penelitian sebelumnya melakukan kegiatan klasterisasi terhadap konten *channel youtube gaming* di Indonesia dengan melakukan analisa terhadap hasil klasterisasi konten *channel youtube gaming* di Indonesia di mana LDA akian merepresentasikan setiap konten menjadi beberapa topik dimana setiap topik tersusun dari distribusi kata-kata. Dalam penelitian sebelumnya ini, dihasilkan beberapa analisa seperti topik yang paling digemari oleh masyarakat, topik yang paling paling banyak digunakan, dan topik yang digemari oleh masyarakat namun kenyataannya kurang dieksploitasi. Penelitian tersebut bermanfaat bagi para konten kreator yaitu untuk menjadi referensi dalam membuat karya selanjutnya karena para konten kreator dapat mengetahui topik dominan yang sering diunggah oleh *youtuber gaming* melalui penelitian tersebut [17].

III. METODOLOGI

Bab ini akan menjelaskan metode penelitian yang dilakukan agar pengerjaan terarah dan sesuai dengan tujuan yang tertulis di latar belakang. Berikut adalah alur metode penelitian.



Gbr 2 Alur Penelitian

A. Studi Literatur

Langkah awal dalam penelitian ini adalah melakukan studi literatur. Studi literatur adalah langkah untuk mencari dan mengumpulkan literatur dari penelitian sebelumnya dengan topik yang sama. Dengan langkah ini kami mencari dan menggali informasi terkait metode yang akan digunakan untuk literatur ini. Informasi yang kami dapatkan dari langkah ini adalah dasar teori mengenai konten pornografi, media sosial Twitter, dan metode LDA.

B. Data Crawling

Langkah ini adalah langkah pengambilan data yang tersedia *online*. Data setelah itu akan diimpor dan diekstraksi lebih lanjut sesuai kebutuhan. Pada penelitian ini, kami melakukan *crawling data* terhadap Twitter. Tahap pertama adalah mengumpulkan cuitan dengan kata kunci porno, sange, dan bokep. Data yang didapatkan dari tahap ini adalah kumpulan kalimat yang mengandung kata kunci tadi. Melalui data ini nantinya akan dicari topik apa yang muncul.

C. Data Preparation

Data yang telah didapat dari langkah sebelumnya kemudian dilakukan proses pembersihan. Proses pembersihan dilakukan dengan menghilangkan tanda baca, karakter, dan beberapa kata yang dirasa tidak perlu. Selain itu, proses ini juga berguna untuk menghilangkan duplikasi, data kosong, dan kesalahan eja. Diharapkan setelah proses ini akan dihasilkan data yang lebih akurat untuk proses selanjutnya.

D. Metode LDA

Langkah selanjutnya adalah metode LDA. Metode ini dapat menemukan topik-topik yang didapat dari data sebelumnya. Metode ini terdiri dari beberapa tahap. Pertama adalah tahap parameter yang akan digunakan sebagai batas data. Tahap selanjutnya adalah semi *random distribution*. Dilanjutkan dengan proses iterasi untuk menentukan distribusi topik.

E. Visualisasi

Data yang telah diproses dengan metode LDA selanjutnya akan menjadi data mentah yang akan dibuat visualisasinya. Visualisasi yang dipilih pada penelitian ini adalah grafik batang. Tujuan dilakukan visualisasi adalah untuk mempermudah analisis data berikutnya. Selain itu visualisasi ini juga dapat membandingkan persebaran data yang satu dengan yang lainnya.

F. Analisis dan Kesimpulan

Pada langkah ini dilakukan analisis akan data yang telah didapat dari proses-proses sebelumnya dan visualisasi data dari proses sebelumnya. Analisis dilakukan dengan *topic modelling*, yaitu mengelompokkan topik tertentu dari data berdasarkan kata kunci tertentu. Dari proses analisis, penulis dapat menarik kesimpulan berupa topik-topik yang mendominasi konten pornografi yang ada pada media sosial Twitter.

IV. HASIL DAN PEMBAHASAN

A. Data Crawling

Dalam penelitian ini, penulis mencari tahu terkait kata kunci yang kerap digunakan di media sosial Twitter dengan mengakses beberapa konten yang mengandung pornografi. Selain itu, penulis juga menggali lebih dalam konten pornografi yang biasanya berada di daftar terkini atau *trend* yang ada di media sosial Twitter. Dengan melakukan pencarian konten yang relevan secara terus menerus, dihasilkan beberapa kata kunci yang seringkali digunakan.

Berikut merupakan kata kunci yang banyak ditemui di konten-konten yang relevan yang selanjutnya akan digunakan pada tahap *data crawling*.

1. Bokep
2. Sange
3. Porno

Dari kata kunci yang ditentukan dari tahapan sebelumnya, dilakukan tahapan selanjutnya yaitu *data crawling*. Tahapan ini dilakukan untuk menemukan cuitan yang mengandung kata kunci sesuai yang ditentukan sebelumnya. *Data crawling* dilakukan dengan menggali cuitan yang relevan di media sosial Twitter yang terunggah dari tanggal 8 September 2022 sampai dengan 15 September 2022. Cuitan yang dipilih dari metode ini adalah cuitan yang berbahasa Indonesia. Dari tahap ini didapatkan dataset yang selanjutnya akan diunggah ke akun Zenodo dan akan diolah untuk tahap selanjutnya [18]. Dari tahapan ini, diketahui banyaknya konten yang mengandung kata kunci sesuai yang ditentukan sebelumnya. Contoh beberapa cuitan dapat dilihat pada gambar tabel berikut.

link	username	tweet	tweet_preprocessed
https://twitter.com/edijuna29896873/status/1570200044900581378	edijuna29896873	@dewirara889 Sange bareng yuk	sange bareng yuk
https://twitter.com/Robocop185/status/1570199421178253312	robocop185	@adawia844 Di remas muluu jadinya sange say😀	remas muluu jadinya sange say
https://twitter.com/BUMxJENNIE/status/1570199297349791744	bumxjennie	@BUMxYUNJIN Pagi ratu sange	pagi ratu sange
https://twitter.com/WitanSulaiman14/status/1570198118293176320	witansulaiman14	@crumeble Aku sange nih	aku sange
https://twitter.com/japrudtst3031/status/1570198106624622592	japrudtst3031	Pagi2 sange call vc yuukk	pagi sange call vc yuukk
https://twitter.com/AgusVijey/status/1570197400006070273	agusvijey	@ArumTal Kalau sange mah ya ngewe ath yayaya	kalau sange mah iya ngewe kalau yayaya
https://twitter.com/abbabsbsbshahah/status/1570197320842747904	abbabsbsbshahah	Brondong smp/sma/kuliah ganteng, kirim foto kalian telanjang dong. Kalau cocok jadi brondong tante 😍 #vcscolmek #vcsbugil #vcs #vcsreal #tantengentot #tantemesum #tantekesepian #SANGE_AAAAAAAAAA #JandaNgentot #jandahot #vcsgratis #vcsgratisan #vcsReady #vcsReady #vcskuy #vcskuy	brondong smpsmakuliah ganteng kirim foto kalian telanjang dong kalau cocok jadi brondong tante vcscolmek vcsbugil vcs vcsreal tantengentot tantemesum tantekesepian sangeaaaaaaaaaa jandangentot jandahot vcsgratis vcsgratisan vcsready vcsready vcskuy vcskuy

Gbr 3 Tabel Hasil Data *Crawling*

B. Data Preparation

Dengan tahapan *data crawling* yang dilakukan sebelumnya dari Twitter, dihasilkan *output* berupa *file* data .csv yang berisi tanggal, waktu, *username*, nama akun, isi cuitan, jumlah *like*, jumlah *mention*, tagar, dan tautan cuitan. Lalu penulis melakukan *data preparation* dengan menyiapkan data yang akan digunakan dalam proses LDA. Dalam proses ini akan digunakan isi cuitan dari kata kunci yang relevan. Total cuitan yang dihasilkan dari proses sebelumnya sebanyak 15.135 cuitan. Dari data yang didapatkan, dilakukan *data cleaning* menggunakan *libabry* Phyton Sastrawi untuk menghapus beberapa isi cuitan yang kurang relevan seperti tagar, url, tanda baca, simbol-simbol, dan emotikon untuk menghasilkan isi cuitan yang telah diproses dengan hanya berisi kalimat.

Hasil *data preprocessing* dilakukan *lower casing* untuk penyamaan *case* huruf dengan mengubah semua karakter menjadi huruf kecil. Lalu dipilih beberapa kolom yang sesuai untuk digunakan pada tahap selanjutnya, di antaranya adalah tautan, isi cuitan, dan *tweet preprocessed* yang hanya mencantumkan isi cuitan yang telah diproses sebelumnya. Setelah itu dilanjutkan dengan *stopword removal* dengan melakukan penghapusan kata-kata yang sering muncul namun tidak memiliki makna, contohnya yaitu: dari, dia, tidak, tak, dan lain sebagainya. Dari proses inipun dapat dihasilkan visualisasi kata kunci yang sering muncul dalam cuitan yang

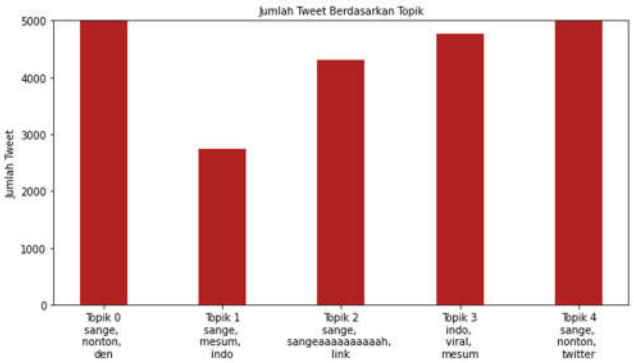
kami dapatkan dalam bentuk *word cloud.* Berikut adalah gambar *word cloud* yang dihasilkan.

Gbr 4 Hasil *Word Cloud*

C. *Metode LDA*

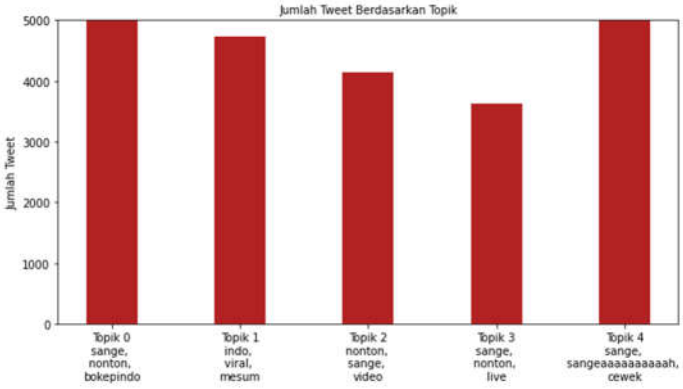
LDA merupakan sebuah metode yang mendeteksi topik pada koleksi dokumen beserta proporsi kemunculan topik tersebut. Ada dua bentuk distribusi probabilitas yang harus dicari dengan LDA, yaitu distribusi probabilitas topik dalam dokumen dan distribusi probabilitas kata dalam topik. Semakin besar nilai *alpha* dalam dokumen, semakin banyak juga topik yang dibahas dalam dokumen tersebut. Lalu untuk distribusi kata dalam topik ditandai dengan pemakaian beta. Semakin kecil beta maka kata-kata yang ada di dalam topik semakin sedikit sehingga topik tersebut mengandung kata-kata yang lebih spesifik. Dalam hal ini, kami menggunakan *alpha* dan *beta* dengan besar yang sama, yaitu 0.01.

Sebelum ke tahap iterasi, terdapat proses pembuatan kamus atau *dictionary* dari semua kata dalam koleksi dokumen. Dalam hal ini dokumen kami adalah kumpulan *tweet prepocessed* yang telah disebutkan pada tahap sebelumnya. Pembuatan kamus ini bermaksud untuk melakukan *indexing* terhadap semua kata dalam dokumen. Karena LDA bersifat *bag-of-words* (BoW) maka informasi urutan antar kata tidak diperhatikan. Setelah membuat kamus dan menentukan *alpha*, *beta*, dan iterasi, maka didapatkan topik dominan dan tiga kata kunci teratas dari tiap topik dominan tersebut. Agar informasi lebih jelas dan lebih mudah dipahami, kami menampilkan hasil LDA dalam bentuk grafik batang. Berikut adalah hasil dari LDA yang kami lakukan dengan 3 kata kunci utama dari setiap topik dominan yang mengandung konten pornografi.



Gbr 5 Hasil LDA

Bisa dilihat dari Gbr **5** Gbr **5** Hasil LDAdi atas, kata kunci yang ada dari 5 topik dominan yang didapatkan. Namun dari gambar di atas dapat dilihat bahwa masih ada kata kunci yang tidak relevan atau tidak mempunyai arti dengan konten pornografi, yaitu 'den'. Oleh karena itu, kami melakukan penambahan kata tersebut ke dalam *stopword* dan melakukan iterasi untuk mendapatkan hasil yang lebih maksimal. Grafik hasil iterasi dapat dilihat pada Gbr **6** di bawah ini.



Gbr 6 Hasil LDA setelah dilakukan *stopword*

Untuk mengevaluasi mode, digunakan dua parameter yaitu nilai *perplexity* dan nilai *coherence*. Nilai *perplexity* menilai tigkat keterkejutan suatu model pada data yang diprosesnya dengan menghitung *normalized log-likehood*. Nilai *perplexity* yang semakin rendah, maka semakin baik model yang dibuat. Sedangkan, nilai *coherence* merupakan nilai yang didapatkan dari perhitungan kemiripan semantik antara kata yang terdapat pada topik. Oleh sebab itu, nilai *coherence* yang semakin tinggi, maka model LDA yang dibuat semakin baik.

D. *Analisis*

Dari hasil percobaan dengan menggunakan metode LDA, didapatkan beberapa hasil dari data tersebut yang dapat dianalisis. Dalam penentuan jumlah topik yang sesuai dengan konten pornografi dapat dilihat dari nilai *perplexity* dan nilai topik *coherence*. Nilai *perplexity* yang semakin rendah maka semakin baik modelnya. Sedangkan nilai *coherence* yang semakin tinggi maka semakin baik modelnya

TABEL I

NILAI PERXPLEXITY DAN NILAI COHERENCE

JUMLAH TOPIK	NILAI PERXPLEXITY	TOPIK COHERENCE
3	-32.463479498133964	0.3876787659275303
4	-32.28188414571735	0.4195310056648073
5	-36.62095757355965	0.4447559528727839

Dari hasil yang didapatkan pada TABEL I**Error! Reference source not found.** ditunjukkan bahwa nilai antar topik memiliki selisih yang cukup besar. Dapat dilihat bahwa nilai *perxplexity* terendah ada pada jumlah topik 5. Begitu pula dengan nilai topik *coherence* tertinggi. Oleh karena itu, terpilihlah jumlah topik dominan sebanyak 5 topik dalam penelitian ini.

Dari grafik yang dihasilkan pada Gbr 6, dapat dilihat bahwa kata kunci 'sange' merupakan kata kunci yang paling banyak digunakan oleh pengguna Twitter untuk menyampaikan atau mengunggah suatu unggahan dengan konten yang mengarah kepada pornografi. Karena kata tersebut muncul pada 4 topik dominan berdasarkan perhitungan LDA. Yang berarti kata tersebut merupakan kata yang paling banyak muncul dalam setiap topik dominan. Secara detail, hasil jumlah dari setiap kata kunci yang sering digunakan dapat dilihat pada tabel di bawah ini.

TABEL II

JUMLAH CUITAN KATA KUNCI DARI TOPIK DOMINAN

TOPIK DOMINAN	KATA KUNCI	JUMLAH CUITAN
0	SANGE, NONTON, BOKEPINDO	7925
1	INDO, VIRAL, MESUM	5142
2	NONTON, SANGE, VIDEO	9096
3	SANGE, NONTON, LIVE	5053
4	SANGE, SANGEAHHH, CEWEK	3687

Berdasarkan TABEL II, topik dominan pertama membahas orang sange yang mencari video atau film bokepindo untuk ditonton. Topik tersebut memiliki jumlah cuitan sebanyak 7925 cuitan. Topik dominan selanjutnya membicarakan video mesum yang viral di Indo (Indonesia). Topik dominan tersebut ditemukan sejumlah 5142 cuitan. Topik dominan yang ketiga membahas mengenai nonton video sange. Topik tersebut mempunyai jumlah yang terbanyak dibandingkan dengan topik dominan lainnya, yaitu dengan 9096 cuitan. Topik dominan keempat membicarakan nonton video *live* (siaran langsung) sange. Topik tersebut memiliki 5053 cuitan. Topik dominan terakhir yaitu cewek sange yang mendesah. Topik itu berjumlah 3687 cuitan.

V. KESIMPULAN

Berdasarkan hasil dan pembahasan yang dilakukan terhadap tiga kata kunci dari Twitter yang dilakukan LDA, maka didapatkan 5 topik dominan dengan 3 kata kunci utama pada setiap topiknya. Pemilihan dengan menampilkan 5 topik ini karena melihat hasil nilai *perxplexity* dan nilai *coherence*.

Dari hasil analisis menggunakan metode LDA yang dilakukan dari konten di media sosial Twitter didapatkan topik dominan teratas yang membahas mengenai menonton video sange. Selanjutnya topik yang menempati urutan kedua mengenai orang sedang memiliki nafsu birahi yang mencari tontonan video bokep Indonesia. Kemudian topik yang menempati urutan ketiga membahas mengenai video mesum yang sedang viral di Indonesia. Untuk topik yang menempati urutan keempat membahas mengenai tontonan siaran langsung orang yang sedang birahi. Selanjutnya, topik yang menempati urutan kelima mengenai perempuan birahi yang sedang mendesah. Berdasarkan data *crawling* yang didapat terdapat beberapa akun yang terdapat cuitan yang sama dengan akun lainnya. Dari beberapa akun tersebut banyak yang menawarkan jasa prostitusi melalui telepon ataupun *video call*.

Berdasarkan hasil LDA yang dilakukan, didapatkan 5 topik dominan dengan tiga kata kunci pada setiap topik dominan yang sering digunakan oleh pengguna Twitter untuk menyebarkan cuitan konten pornografi. Hasil tersebut diharapkan dapat membantu Kominfo untuk dapat menindaklanjuti pelaku yang menyebarluaskan konten pornografi di media sosial Twitter. Tindakan penindak lanjutan yang dapat dilakukan oleh Kominfo dengan hasil ini seperti melaporkan kepada perusahaan Twitter terkait kata kunci yang sering digunakan yang selanjutnya dapat menemukan pelaku penyebarluasan konten pornografi dan memblokir akun tersebut dan melaporkan pelaku-pelaku selanjutnya yang menggunakan kata kunci tersebut.

Metode pengumpulan data dengan menggunakan kata kunci memiliki keterbatasan karena pada penelitian ini hanya menggunakan beberapa kata kunci dalam pencarian. Sehingga mungkin hanya menampilkan gambaran yang kurang utuh dari penyebaran konten pornografi di Twitter. Pada penelitian berikutnya dapat menggunakan beberapa kata kunci yang lebih representatif dalam penyebaran konten pornografi dan menggunakan metode LDA yang lebih efektif dengan langsung mengimpor seluruh *file* tanpa terbatas pada durasi waktu tertentu.

REFERENSI

[1] I. G. P. Udayana, I. M. M. Widyantara, and N. M. S. Karma, "Penyalahgunaan Aplikasi Media sosial sebagai Eksploitasi dalam Tindak Pidana Pornografi," *Jurnal Konstruksi Hukum*, vol. 3, no. 2, pp. 438–443, Mar. 2022, doi: 10.55637/JKH.3.2.4852.438-443.

[2] "Digital 2021: the latest insights into the 'state of digital' - We Are Social UK," Jan. 27, 2021.

https://wearesocial.com/uk/blog/2021/01/digital-2021-the-latest-insights-into-the-state-of-digital/ (accessed Oct. 02, 2022).

[3] Kementerian Komunikasi dan Informatika, "Ada 431.065 Aduan Konten Negatif, Mayoritas Pornografi," Jan. 09, 2020. https://www.kominfo.go.id/content/detail/23717/ada-431065-aduan-konten-negatif-mayoritas-pornografi/0/sorotan_media (accessed Oct. 02, 2022).

[4] "Lindungi Anak dari Bahaya Pornografi Online | Indonesia Baik." https://indonesiabaik.id/infografis/lindungi-anak-dari-bahaya-pornografi-online (accessed Oct. 02, 2022).

[5] "UNDANG-UNDANG REPUBLIK INDONESIA NOMOR 44 TAHUN 2008 TENTANG PORNOGRAFI".

[6] M. Andriyani and M. Ardina, "Pengaruh Paparan Tayangan Pornografi melalui Media Sosial terhadap Perilaku Mahasiswa di Yogyakarta," *Jurnal Audiens*, vol. 2, no. 1, pp. 143–153, Mar. 2021, doi: 10.18196/jas.v2i1.11138.

[7] M. Taufiq Anwar, A. Iriani, D. Herman, F. Manongga, and K. Satya Wacana, "Analisis Pola Persebaran Pornografi pada Media Sosial dengan Social Network Analysis," *Jurnal Buana Informatika*, vol. 9, no. 1, pp. 43–52, Jul. 2018, doi: 10.24002/JBI.V9I1.1667.

[8] D. Eka Saputra, J. Adhyaksa No, and K. Banjarmasin Kalimantan Selatan, "KAJIAN YURIDIS TERHADAP TINDAK PIDANA PORNOGRAFI MELALUI MEDIA SOSIAL," *Al-Adl : Jurnal Hukum*, vol. 9, no. 2, pp. 263–286, Nov. 2017, doi: 10.31602/AL-ADL.V9I2.949.

[9] A. M. Pudjajana and D. Manongga, "Sentimen Analisis Tweet Pornografi Kaum Homoseksual Indonesia di Twitter dengan Naive Bayes ," *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, vol. 9, no. 1, pp. 313–318, Apr. 2018, doi: 10.24176/SIMET.V9I1.1922.

[10] D. Gunawan, R. Mahardika, F. Ranja, S. Purnamawati, and I. Jaya, "The Identification of Pornographic Sentences in Bahasa Indonesia," *Procedia Comput Sci*, vol. 161, pp. 601–606, Jan. 2019, doi: 10.1016/J.PROCS.2019.11.162.

[11] M. A. N. Febriansyach, F. Rashif, G. I. P. Nirvana, and N. A. Rakhmawati, "Implementasi LDA untuk Pengelompokan Topik Tweet Akun Bot Twitter bertagar #covid-19," *CogITo Smart Journal*, vol. 7, no. 1, p. 170, Jun. 2021, doi: 10.31154/COGITO.V7I1.299.170-181.

[12] A. Rahmawati, N. L. Nikmah, R. D. A. Perwira, and N. A. Rakhmawati, "Analisis topik konten channel YouTube K-pop Indonesia menggunakan Latent Dirichlet Allocation," *Teknologi*, vol. 11, no. 1, pp. 16–25, Jan. 2021, doi: 10.26594/TEKNOLOGI.V11I1.2155.

[13] A. I. Prakerti, A. F. Claresta, M. R. K. Ibrahim, and N. A. Rakhmawati, "Model Latent Dirichlet Allocation Pada Perilaku Siswa Menggunakan Media Pembelajaran Daring," *INFORMATION MANAGEMENT FOR EDUCATORS AND PROFESSIONALS : Journal of Information Management*, vol. 5, no. 1, pp. 35–44, Dec. 2020, doi: 10.51211/IMBI.V5I1.1407.

[14] K. R. Adjie, P. Santoso, A. Husna, N. W. Putri, and A. Rakhmawati, "Analisis Topik Tagar Covidindonesia pada Instagram Menggunakan Latent Dirichlet Allocation," *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 7, no. 1, pp. 1–9, Jan. 2022, doi: 10.14421/jiska.2022.7.1.1-9.

[15] F. Z. Ahmad, M. F. S. Arifandy, M. R. Caesarardhi, and N. A. Rakhmawati, "Bagaimana Masyarakat Menyikapi Pembelajaran Tatap Muka: Analisis Komentar Masyarakat pada Media Sosial Youtube Menggunakan Algoritma Deep Learning Sekuensial dan LDA," *Jurnal Linguistik Komputasional*, vol. 4, no. 2, pp. 40–46, Nov. 2021, doi: 10.26418/JLK.V4I2.57.

[16] Y. Sahria and D. Hatta Fudholi, "Analysis of Health Research Topics in Indonesia Using the LDA (Latent Dirichlet Allocation) Topic Modeling Method," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 2, pp. 336–344, Apr. 2020, doi: 10.29207/RESTI.V4I2.1821.

[17] D. A. Rahman, R. B. Waskitho, M. Fajrul, A. U. Nuha, and N. A. Rakhmawati, "Klasterisasi Topik Konten Channel Youtube Gaming Indonesia Menggunakan Latent Dirichlet Allocation," JIEET (Journal of Information Engineering and Educational Technology), vol. 5, no. 2, pp. 78–83, Dec. 2021, doi: 10.26740/JIEET.V5N2.P78-83.

[18] V. R. Dianutami, Q. N. Muna, R. A. Zuhroh, and N. A. Rakhmawati, "LDA-TwitterPornografi: Codingan Klasterisasi dengan LDA Twitter," Oct. 2022, doi: 10.5281/ZENODO.7134514.