Analitik Big Data: Social Media Mining

Annisa Rahmala[1], Don Ardhito[2], Leo Riska[3]
[1,2,3]Program Studi Magister Ilmu Komputer Universitas Budi Luhur
Email : [*1]**1911601456@student.budiluhur.ac.id**, [2]1911601449@student.budiluhur.ac.id,
[3]1911600920@student.budiluhur.ac.id

Abstrak

Perkembangan big data dari tahun ke tahun menjadi sangat pesat sejalan dengan perekembangan Internet of Things (IoT). Salah satu device yang paling banyak dimiliiki dan sering digunakan adalah mobile phone terutama smartphone. Aplikasi media sosial yang terdapat pada smartphone menghasilkan data dalam jumlah besar setiap detiknya. Data-data tersebut banyak digunakan untuk penelitian diberbagai bidang analitik big data. Dan teknik yang digunakan juga beragam tergantung pada dataset dan tujuan dilakukannya penelitian. Penelitian ini bertujuan untuk membahas tentang big data analitik khususnya social media mining, dan beberapa tekniknya untuk anaisis. Penelitian dilakukan dengan metode kajian literatur terhadap jurnal dan buku membahas topik yang relevan dengan penelitian ini. Manfaat yang diharapkan dari penelitian ini adalah dapat menambah wawasan dan pengetahuan bagi pembaca yang tertarik dengan topik social media mining.

Kata Kunsi: analitik big data, social media mining, data mining

Abstract

The development of big data from year to year has become very rapid in line with the development of the Internet of Things (IoT). One of the most widely owned and frequently used devices is a mobile phone, especially a smartphone. Social media applications found on smartphones generate large amounts of data every second. These data are widely used for research in various fields of big data analytics. And the techniques used also vary depending on the dataset and the purpose of the research. This study aims to discuss big data analytics, especially social media mining, and some of its techniques for analysis. The research was conducted using a literature review method on journals and books discussing topics relevant to this research. The expected benefit of this research is that it can add insight and knowledge for readers who are interested in the topic of social media mining.

Keywords: *big data analytics, social media mining, data mining*

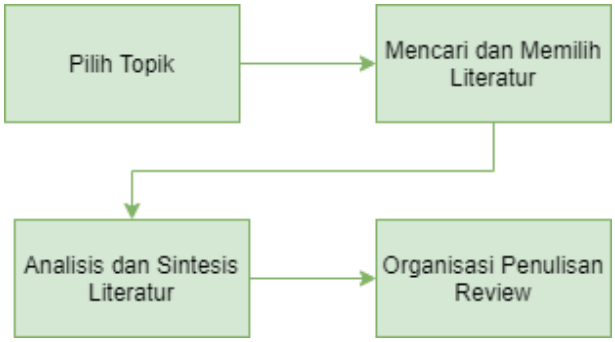
1. PENDAHULUAN

Big data menjadi booming sejak 2011 dan perkembanganya semakin pesat dari tahun ke tahun [1]. Big data yang menjadi hype saat ini adalah big data yang terkait dengan digitalisasi dan penggunaan Internet of things (IoT) yang memproduksi data dengan jumlah yang sangat besar setiap detiknya. Ada berbagai macam device IoT yang digunakan saat ini, device yang paling sering digunakan adalah mobile phone (smartphone), tablet, personal computer, laptop, notebook dan lain sebagainya. Dalam penelitian disebutkan sekitar 3 miliar orang memiliki mobile phones, dan personal computer yang menghasilkan miliaran data per detik dalam bentuk teks, gambar, suara, dan video. Semua data yang dihasilkan tersebut kebanyakan berasal dari aplikasi media sosial seperti We Chat, Twitter, Facebook dan Instargram yang memiliki jumlah pengguna yang sangat besar dan tersebar diseluruh dunia. Aplikasi *social media* telah menjadi bagian dari kehidupan sehari-hari dari orang-orang yang secara terus-menerus berbagi pendapat mereka tentang hidup, informasi, pengetahuan, minat dan sebagainya dalam setiap detiknya [3].

Data dari media sosial menjadi sumber penemuan pengetahuan terbesar dan banyak digunakan bidang analitik big data dengan berbagai metode dan teknik [4]. Flickr digunakan untuk mengklasifikasi tutupan lahan dengan menggunakan artificial neural networks (ANNs) [5]. Analisis peringatan dini bencana alam dan penilaian kerusakan akibat bencana dapat menggunakan dari data media sosial dan informasi geo-location dengan melakukan metode sentiment analysis, dan top hashtags and keywords frequency analysis [6]. Teks dan data mining dari media sosial digunakan untuk membuat mapping aktivitas rekreasi satwa liar [7]. Social media data mining digunakan untuk mendeteksi perilaku teoristik melalui karakteristik Tweet berbahasa Arab [8].

Artikel ini bertujuan untuk membahas hasil review dari beberapa penelitian analitik big data dalam bidang social media data mining. Pada bagian kedua akan dibahas mengenai metodologi penelitian. Bagian ketiga membahas hasil dan diskusi penelitian yang berisi tentang penggunaan medial sosial dalam analitik big data, teknik social media mining, dan komparasi beberapa metode social media mining. Bagian terakhir membahas kesimpulan yang didapatkan dari penelitian.

2. METODE PENELITIAN



Gambar 2.1 Metode Penelitian

Metodologi penelitian yang digunakan dalam penelitian ini menggunakan 4 tahapan penelitian yang dapat dilihat pada gambar 2.1, yaitu pilih topik, mencari dan memilih literature, analisis dan sintesis literature, dan organisasi penulisan review. Berikut ini adalah penjelasan tahapannya: (1) Pemilihan topik, yaitu menentukan topik untuk dijadikan bahan pembahasan dan review; (2) Studi literatur, literatur yang dicari dan dipilih adalah literatur yang relevan dan sesuai dengan topik yang dipilih; (3) Analisis dan sintesis literatur, dilakukan analisis dan sinstesis literatur yang menjadi poin penting dari penelitian yang dilakukan; (4) Organisasi penulisan review, membuat kerangka penulisan review dan kemudian memulai penulisan.

3. HASIL DAN PEMBAHASAN

3.1. Penggunaan Media Sosial Dalam Analitik Big Data

Tabel 3.1 Penelitian-Penelitian Big Data yang Menggunakan Media Sosial

No.	Penulis	Tujuan	Metode
1	Hu, dkk. [9]	Mengeksplorasi local users's sentiment yang diekstrak dari data Geo-tweets data dari Januari hingga Desember 2016, dan	– Valence Aware Dictionary and sEntiment Reasoner (VADER) – Local Indicators of Spatial

		dianalisis dalam prespektif spasial dan temporal	Association (LISA) – Latent Dirichlet Allocation (LDA) model
2	Kolahkaj, dkk.[10]	Menghasilkan rekomendasi paket perjalanan yang dinamis dengan mempertimbangkan data multidimensi waktu, lokasi, peringkat implisit pengguna, karakteristik wisatawan dan pola pergerakan perjalanan berdasarkan geo-tagged photos.	– Algoritma PrefixSpan – Algoritma DBSCAN
3	Park, dkk.[11]	Menganalisis dataset mobile phone dalam skala besar yang didapatkan dari jejak ponsel wisatawan internasional yang mengunjungi Korea Selatan	- Algoritma DBSCAN - Algoritma SPADE
4.	ElQadi, dkk.[5]	Menghasilkan framework yang dapat digunakan kembali untuk menemukan dan memfilter citra (imaginary) untuk menemkukan jenis tutupan lahan.	Artificial neural networks
5	Desheng Wu, dan Yiwen Cui [6]	Mengembangkan framework yang mengabungkan data dari berbagai sumber termasuk media sosial (Twitter), dan lokasi geografis, dengan informasi tentang kerugian akibat bencana	SentiWordNet 3.0
6	Yan, dkk.[12]	Memanfaatkan data media sosial geo-tagged untuk menilai pemulihan pasca bencana tempat pariwisata melalui penambangan sentiment dan perspektif orang-orang tentang status pemulihan.	- PatternAnalyzer - Latent Dirichlet allocation (LDA)
7	Hou, dkk.[13]	Eksplorasi public attention pada media sosial terkait dengan COVID-19	- Latent Dirichlet Allocation (LDA) - Baidu Sentiment Analysis
8	Monkman, dkk [7]	Menguraikan pendekatan yang dapat diakses untuk penggunaan teks dan data mining media sosial untuk mengumpulkan informasi tentang kegiatan rekreasi satwa liar.	text and data mining (TDM), untuk klasifikasinya menggunakan leksikon yang didapatkan dari pengetahuan ahli dan ulasan forum diskusi.
9	Alhalabi, dkk.[8]	Mengusulkan sistem yang menyediakan algoritma cerdas untuk mendeteksi perilaku teroris dengan menggunakan karakteristik dinamis dari informasi dan aktivitas yang dibagikan di media sosial	Algoritma yang diusulkan oleh penulis
10	Abdul-Rahman, dkk.[14]	Mengusulkan framework untuk penambangan big data media sosial dan analitik data	- Latent Dirichlet Allocation (LDA) - VADER (Valence Aware

		menggunakan Twitter untuk perencanaan dan pengolahan kota yang berkelanjutan	Dictionary for sEntiment Reasoning)

Berdasarkan pada tabel 3.1 media sosial dapat digunakan diberbagai bidang penelitian yang berkaitan dengan big data, misalnya dalam bidang kesehatan [15], [13]; pembangunan kota [14], [5], [9] ; pariwisata [11], [10][12]; sosial [8]; bencana alam [5], [16]; biological conservation [7]. Semua penelitian tersebut digunakan untuk membantu dalam pengambilan keputusan. Metode yang digunakan juga beragam, ada yang mengusulkan metode baru yang dibuat sesuai dengan kebutuhan penelitian [8], dan ada juga yang menggunakan metode yang sudah tersedia. Pembahasan mengenai metode dalam sosial media mining akan dibahas beberapa di sub-bab selanjutnya.

3.2. Teknik Sosial Media Mining

Data mining adalah sebuah tahapan untuk melakukan ekstraksi informasi yang memiliki kegunaan dan relasi secara otomatis dari data yang kuantitasnya besar sekali [17]. Dengan menambang data, dimana merupakan sebuah ilmu yang baru kita akan mampu untuk melakukan ekstraksi terhadap informasi yang dimiliki dari data tersebut. Data mining mengungkapkan akan adanya relasi yang memiliki manfaat antar data, dan hal ini memungkinkan untuk diterapkan dengan tujuan untuk menciptakan sebuah *decision making* yang tepat [20]. Data mining dalam penggunaannya juga didukung oleh knowledge lainnya seperti neural network, image database, pattern recognition, signal processing, spatial data analysis [18]. Data mining memiliki karakteristik seperti [18]: (1) *Discovery*, dimana dengan melakukan penambangan data, kita dapat menemukan suatu informasi yang tersembunyi dan *pattern* tertentu yang sebelumnya belum diketahui; (2) Penambangan data pada umumnya bekerja dengan data dengan kuantitas yang sangat besar dengan tujuan agar hasilnya lebih dapat dipercaya; (3) Penambangan data memiliki manfaat pengambilan keputusan yang lebih kritis, khususnya dalam strategi. Dengan media sosial data dapat dihasilkan serta disebarluaskan di domain publik dan jenis data ini menarik bagi banyak pemangku kepentingan dan penerima manfaat dalam berbagai sektor. Hal ini dipengaruhi oleh isi dari konten yang tidak terstruktur dan tidak disensor yang dapat memicu sentimen yang berbeda dari pengguna ke masyarakat luas dan menghasilkan suatu informasi. Komunikasi di media sosial dilakukan secara real time dan aspek komunikasi yang tepat waktu dan ketepatan dalam pelaporan sangat penting di bidang-bidang seperti kesehatan, pariwisata, keamanan dan pendidikan [21].

Dalam melakukan penambangan data, baik itu dari media sosial, instansi tertentu atau media lainnya, ada 6 tahapan yang harus dilakukan yaitu [17]: (1) Pembersihan data, (2) Integrasi; (3) Pemilihan data, (4) Transformasi data; (5) Penambangan; (6) Evaluasi.

Data mining dalam pengaplikasiannya merupakan salah satu bagian dari KDD (Knowledge Discovery in Database) dimana bertujuan untuk melakukan ekstraksi pattern dan model dari data dengan menggunakan suatu algoritma. Sedangkan Knowledge Discovery in Database sendiri merupakan proses penggunaan database dalam jumlah yang besar seperti seleksi, preprocessing, transform, mining, dan evaluasi pola yang terbentuk sehingga dapat diproses menjadi informasi yang bermanfaat. Beberapa algoritma yang sering digunakan untuk melakukan data mining seperti Naïve Bayes, K-NN, Adaboost, dan random forest.

Naïve Bayes merupakan algoritma pengklasifikasi yang paling sederhana dan paling banyak digunakan dan bahkan dapat digunakan untuk kumpulan data yang kecil [20]. Jika model Naïve Bayes standar digunakan untuk mengklasifikasi tweets dalam jumlah yang besar ($\geq$20.000 tweet) hasil akurasi dan presisinya relative rendah sehingga perlu dikombinasikan dengan algoritma yang lain untuk menutup kelemahan tersebut [21].

Random forest merupakan teknik pembelajaran ensembel yang menggunakan beberapa pengklasifikasi pohon keputusan untuk menentukan label untuk titik baru. Teknik ini menentukan label untuk node baru berdasarkan bobot dari setiap pohon keputusan. Random forest dibuat dengan harapan dapat meningkatkan akurasi dari teknik Naïve Bayes dan K-NN,

tetapi pengklasifikasi mendapat perolehan terburuk dibandingkan dengan Naïve Bayes dan K-NN [21]. Jadi random forest juga perlu dikombinasikan dengan algoritma atau metode lain untuk mendapatkan hasil akurasi yang tinggi.

K-Nearest Neighbour (K-NN) menentukan titik k-terdekat ke titik yang saat ini sedang dipertimbangkan. Setelah poin-poin teridentifikasi, pengkalsifikasi menerapkan teknik pemungutan suara mayoritas untuk menentukan label poin baru. Dalam hal presisi K-Nearest Neighbor memiliki performa di atas Naïve Bayes, tetapi menghasilkan recall yang lebih rendah, dan jika jumlah fiturnya banyak tingkat akurasinya akan meningkat [21].

Setiap iterasi pengklasifikasi pada Adaboost meningkatkan akurasi model sebelumnya dengan membobot ulang titik yang salah diklasifikasikan untuk menggeser model baru kea rah meminimalkan total eror. Pengklasfikasi dilakukan menggunakan Adaboost dengan pemrosesan teks dapat meningkatkan nilai akurasi dibandingkan dengan Naïve Bayes, K-NN, dan random forest. Adaboost dapat meningkatkan akurasi karena sifat ansambel yang meningkatkan model pembelajaran secara iteratif [21].

3.3. Perbandingan Metode Social Media Mining

Tabel 3.2 Perbandingan Metode Social Media Mining

Judul Penelitian	Metode	Accuracy
Supervised Learning for Suicidal Ideation Detection in Online User Content [23]	Random Forest and Different Features	96.38%
Text mining and determinants of sentiments: Twitter social media usage by traditional media houses in Uganda [24]	Random Forest and Logistic Regression Model	72%
Analisis Sentimen Data Twitter Menggunakan Metode Text Mining Tentang Masalah Obesitas di Indonesia [25]	Naive Bayes Classifier	94%

Berdasarkan tabel 3.2 adanya penggunaan metode yang sama yaitu algoritma Random Forest dimana algoritma tersebut digunakan pada penelitian lain dan menghasilkan selisih tingkat keakuratan yang cukup jauh. Dapat dilihat bahwa penggabungan metode dan fitur tertentu dapat menghasilkan tingkat akurasi yang cukup tinggi dibandingkan metode Random Forest yang digabungkan dengan model Logistic Regression menghasilkan tingkat akurasi yang rendah.

4. KESIMPULAN

Platform media sosial memberi dampak yang sangat besar terhadap informasi yang dihasilkan dari berbagai kondisi dan sector. Studi tentang penambangan data dari media sosial yaitu mengekstrak pengetahuan dari kumpulan data tekstual yang sangat besar. Data dari sosial media dapat digunakan dalam berbagai bidang analitik big data untuk membantu dalam pengambilan keputusan. Terdapat berbagai jenis metode atau teknik yang digunakan untuk analitik big data. Metode tersebut memiliki kelebihan dan kelemahan tersendiri. Kelemahan dari

metode-metode tersebut perlu ditutupi dengan mengkombinasikanya dengan metode, model, dan atau teknik yang lain. Hasil perbandingan juga menunjukkan untuk dapat meningkatkan tingkat akurasi, metode perlu dikombinasikan dengan metode yang lain.

5. SARAN

Disarankan agar analitik big data: social media mining ini yang telah dibuat ini dapat membantu di dalam pengambilan keputusan yang tepat dan bermanfaat bagi para pengguna media social.

DAFTAR PUSTAKA

[1] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, Apr. 2015, doi: 10.1016/j.ijinfomgt.2014.10.007.

[2] M. A. Carlos, M. Nogueira, and R. J. Machado, "Analysis of dengue outbreaks using big data analytics and social networks," in *2017 4th International Conference on Systems and Informatics, ICSAI 2017*, 2017, vol. 2018-January, pp. 1592–1597, doi: 10.1109/ICSAI.2017.8248538.

[3] "Sentiment Analysis On Social Media Big Data With Multiple Tweet Words," doi: 10.35940/ijitee.J9684.0881019.

[4] M. M. ElQadi, M. Lesiv, A. G. Dyer, and A. Dorin, "Computer vision-enhanced selection of geo-tagged photos on social network sites for land cover classification," *Environ. Model. Softw.*, vol. 128, no. March, p. 104696, 2020, doi: 10.1016/j.envsoft.2020.104696.

[5] D. Wu and Y. Cui, "Disaster early warning and damage assessment analysis using social media data and geo-location information," *Decis. Support Syst.*, vol. 111, no. October 2017, pp. 48–59, 2018, doi: 10.1016/j.dss.2018.04.005.

[6] G. G. Monkman, M. J. Kaiser, and K. Hyder, "Text and data mining of social media to map wildlife recreation activity," *Biol. Conserv.*, vol. 228, no. September, pp. 89–99, 2018, doi: 10.1016/j.biocon.2018.10.010.

[7] W. Alhalabi *et al.*, "Social mining for terroristic behavior detection through Arabic tweets characterization," *Futur. Gener. Comput. Syst.*, vol. 116, pp. 132–144, 2021, doi: 10.1016/j.future.2020.10.027.

[8] T. Hu, B. She, L. Duan, H. Yue, and J. Clunis, "A systematic spatial and temporal sentiment analysis on geo-tweets," *IEEE Access*, vol. 8, pp. 8658–8667, 2020, doi: 10.1109/ACCESS.2019.2961100.

[9] M. Kolahkaj, A. Harounabadi, A. Nikravanshalmani, and R. Chinipardaz, "A hybrid context-aware approach for e-tourism package recommendation based on asymmetric similarity measurement and sequential pattern mining," *Electron. Commer. Res. Appl.*, vol. 42, no. February, p. 100978, 2020, doi: 10.1016/j.elerap.2020.100978.

[10] S. Park, Y. Xu, L. Jiang, Z. Chen, and S. Huang, "Spatial structures of tourism destinations: A trajectory data mining approach leveraging mobile big data," *Ann. Tour. Res.*, vol. 84, no. January, p. 102973, 2020, doi: 10.1016/j.annals.2020.102973.

[11] Y. Yan, J. Chen, and Z. Wang, "Mining public sentiments and perspectives from geotagged social media data for appraising the post-earthquake recovery of tourism destinations," *Appl. Geogr.*, vol. 123, no. February, p. 102306, 2020, doi: 10.1016/j.apgeog.2020.102306.

[12] K. Hou, T. Hou, and L. Cai, "Public attention about COVID-19 on social media: An investigation based on data mining and text analysis," *Pers. Individ. Dif.*, vol. 175, no. September 2020, p. 110701, 2021, doi: 10.1016/j.paid.2021.110701.

[13] M. Abdul-Rahman, E. H. W. Chan, M. S. Wong, V. E. Irekponor, and M. O. Abdul-Rahman, "A framework to simplify pre-processing location-based social media big data

for sustainable urban planning and management," *Cities*, vol. 109, no. September, p. 102986, 2021, doi: 10.1016/j.cities.2020.102986.

[14] K. G. Blumenthal *et al.*, "Mining social media data to assess the risk of skin and soft tissue infections from allergen immunotherapy," *J. Allergy Clin. Immunol.*, vol. 144, no. 1, pp. 129–134, Jul. 2019, doi: 10.1016/j.jaci.2019.01.029.

[15] J. A. de Bruijn, H. de Moel, B. Jongman, J. Wagemaker, and J. C. J. H. Aerts, "TAGGS: Grouping Tweets to Improve Global Geoparsing for Disaster Response," *J. Geovisualization Spat. Anal.*, vol. 2, no. 1, 2018, doi: 10.1007/s41651-017-0010-6.

[16] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition*, vol. 9780470908. 2014.

[17] "Data Mining Concepts and Techniques (2nd Edition) | Request PDF." [Online]. Available: https://www.researchgate.net/publication/262562891_Data_Mining_Concepts_and_Techniques_2nd_Edition. [Accessed: 22-May-2021].

[18] M. Allahyari *et al.*, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," *arXiv*, vol. 13, Jul. 2017.

[19] N. Siyam, O. Alqaryouti, and S. Abdallah, "Mining government tweets to identify and predict citizens engagement," *Technol. Soc.*, vol. 60, p. 101211, 2020, doi: 10.1016/j.techsoc.2019.101211.

[20] S. Ji, C. P. Yu, S. F. Fung, S. Pan, and G. Long, "Supervised learning for suicidal ideation detection in online user content," *Complexity*, vol. 2018, 2018, doi: 10.1155/2018/6157249.

[21] F. Namugera, R. Wesonga, and P. Jehopio, "Text mining and determinants of sentiments: Twitter social media usage by traditional media houses in Uganda," *Comput. Soc. Networks*, vol. 6, no. 1, 2019, doi: 10.1186/s40649-019-0063-4.

[22] F. Fridom Mailo *et al.*, "Analisis Sentimen Data Twitter Menggunakan Metode Text Mining Tentang Masalah Obesitas di Indonesia," *J. Sist. Inf. Kesehat. Masy. J. Inf. Syst. Public Heal.*, vol. 4, no. 1, pp. 28–36, 2019.