Analisis sentimen dan pemodelan topik pandemi Covid-19 pada media sosial Twitter menggunakan Naïve Bayes Classifier dan Latent Dirichlet Allocation.

Herjuna Ardi Prakosa*[1], Ari Budi Riyanto[2] dan Siti Nasiroh[3]

1 **Fakultas Sains dan Teknik Universitas Perwira Purbalingga, Indonesia**
herjuna@unperba.ac.id
2 **Fakultas Sains dan Teknik Universitas Perwira Purbalingga, Indonesia**
aribudiriyanto@unperba.ac.id
3 **Fakultas Sains dan Teknik Universitas Perwira Purbalingga, Indonesia**
sitinasiroh@unperba.ac.id

Abstrak

Virus Corona atau Covid-19 menjadi perhatian khusus diseluruh dunia. Banyak masyarakat membicarakan virus ini melalui unggahan komentar dan opini di media sosial. Twitter merupakan salah satu media sosial yang saat ini masih banyak digunakan masyarakat untuk menyampaikan opini berupa kumpulan kata atau pembicaraan yang disebut *tweets*. Pembicaraan di twitter yang berkaitan dengan topik covid-19 ini dapat di klasifikasikan menggunakan metode pemodelan topik (*topic modelling*)untuk menghasilkan sebuah data topic yang sering dibicarakan pengguna twitter. Salah satu algoritma yang digunakan untuk melakukan pemodelan topik adalah menggunakan *latent dirichlet alocation* (LDA). Pada penelitian ini, LDA digunakan untuk mengetahui kata-kata apa saja yang banyak muncul pada data *tweets* tentang Covid-19 yang telah di unggah masyarakat melalui twitter. Sebelum data *tweets* dimodelkan dengan LDA, dilakukan terlebih dahulu analisis sentimen dengan *naïve bayes classifier* untuk menghasilkan sentimen positif, negatif dan netral. Sebanyak 5000 *tweets* digunakan sebagai dataset untuk diklasifikasikan dengan pemodelan topik. Semua *tweets* yang ditambang masih perlu dilakukan *preprocessing* teks yang bertujuan untuk menghapus kata-kata yang tidak baku, menghapus tanda baca, dan menghapus kata penyambung. *Tweets* yang sudah dilakukan *preprocessing* teks lalu diberikan nilai bobot sehingga diketahui kata apa saja yang banyak muncul dalam tweets yang berkaitan dengan Covid-19. Kata-kata yang banyak muncul dan sudah diberikan bobot akan divisualisasikan menggunakan *world cloud* sehingga dapat dilihat pemetaan kata apa saja yang banyak muncul dalam bentuk gambar.

Kata Kunci pemodelan topik, LDA, NBC, covid-19, klasifikasi

1 Pendahuluan

Virus Covid-19 dideklarasikan oleh *World Health Organization* (WHO) sebagai pandemi global pada 11 Maret 2020 [1]. WHO melaporkan lebih 52 juta orang terkonfirmasi positif Covid-19 dan 1,2 juta orang meninggal dunia pada minggu kedua bulan November 2020. Sementara Indonesia mencatat 463 ribu orang terkonfirmasi positif dengan korban meninggal

* Corresponding author.

telah mencapai 15.148 orang[1]. Covid-19 merupakan penyakit menular yang disebabkan oleh SARS-CoV-2. Dilihat dari cepatnya peningkatan kasus pasien positif COVID-19 di sentimen negatif yang terkesan mempolitisasi dalam pemberitaan, dikeluarkannya kebijakan – kebijakan pemerintah yang dinilai masyarakat terlalu santai dan masih kurang tanggap dalam penanganan penyebaran virus, tidak transparannya pemberitaan saat awal kasus terjadi. Banyaknya pernyataan dari tokoh publik yang dinilai tidak pantas dalam kondisi ini, juga tidak tertibnya masyarakat terhadap kebijakan yang sudah ditetapkan permerintah terkait social distancing sehingga menimbulkan semakin banyaknya kasus terjadi dan kepanikan masyarakat yang semakin menjadi–jadi [1]. Tidak hanya sentimen dan topik negatif saja yang beredar dimasyarakat, tetapi juga ada sentimen positif yang disampaikan masyarakat melalui berbagai macam media, untuk memberikan dukungan kepada pemerintah. Satuan tugas Covid-19, dan tenaga medis di seluruh Indonesia, serta muncul juga pro dan kontra terhadap vaksin juga banyak muncul yang dibicaraan masyarakat. Sentimen dan topik yang ada dimasyarakat banyak disampaikan melalui media sosial. Saat ini salah satu media sosial yang menjadi parameter topik pembicaraan masyarakat pengguna media sosial salah satunya adalah twitter [2].

Analisis sentimen positif, netral, dan negatif untuk mengetahui sentimen masyarakat pengguna twitter tentang topik kesehatan pada masa pandemi Covid-19 dapat mengadopsi metode klasifikasi yaitu naïve bayes classifier (NBC). Metode ini merupakan sebuah metode klasifikasi yang memanfaatkan sebuah nilai dari probabilitas statistika sederhana yang mengasumsikan independen yang kuat dari aturan Bayesian [3]. Metode analis Bayesian ini melakukan sebuah analisis berdasarkan sebuah informasi prior dan informasi sampel. Gabungan dari informasi prior dengan informasi sampel tersebut diberi nama peluang posterior [4].

Pemodelan topik merupakan pengelompokan data teks berdasarkan suatu topik tertentu. Salah satu metode dalam pemodelan topik adalah *latent dirichlet allocation* (LDA) [5]. Metode ini merupakan unsupervised learning aatau tidak membutuhkan data berlabel. Pemodelan topik bekerja seperti *clustering* dengan mengelompokan dokumen berdasarkan kemiripannya. Pada penelitian ini algoritma LDA digunakan untuk mengetahui tren topik di twitter yang berkaitan dengan kesehatan agar dapat menghasilkan sebuah topik kesehatan yang dapat di identifikasi dimasa pandemi covid-19 [6].

Analisis sentimen opini publik Bahasa Indonesia terhadap wisata taman mini Indonesia indah menggunkan NBC dan *particle swarm optimization* didapati akurasi sebesar 70% dengan menggunakan 4 *fold cross validation*[3]. Penelitian Sudiantoro dan Zuliarso, 2018 menyebutkan bahwa algoritma NBC sangat efektif untuk digunakan sebagai proses klasifikasi tweet yang dibutuhkan dalam sistem analisis sentimen ini dimana nilai yang di dapatkan dalam pengujian sampai 84%. Metode NBC dapat digunakan untuk melakukan klasifikasi tweets dengan cukup baik pada sistem analisis sentimen [7].

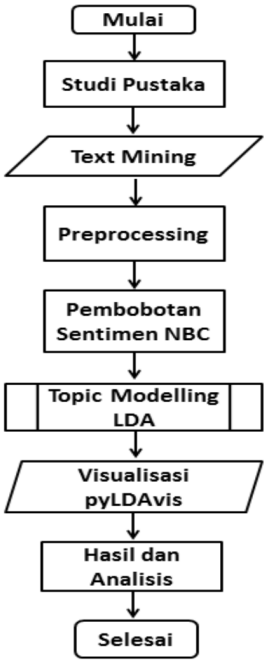
Penelitian membangun model untuk melakukan klasifikasi tweet berdasarkan sentimen dan kategori dengan NBC didapatkan hasil akurasi pengujian klasifikasi dengan fitur *term frequency* diperoleh sebesar 79,91%. Sedangkan fitur TF-IDF didapatkan akurasi sebesar 79,68%. Klasifikasi menggunakan tools RapidMiner dengan Naive Bayes dan fitur term frequency diperoleh sebesar 73,81% sedangkan dengan fitur TF-IDF diperoleh sebesar 71.11% [8].

Pemodelan topik menggunakan metode LDA terbukti dapat melakukan pemodelan topik terhadap judul penelitian di bidang penelitian kesehatan di Indonesia. Hasil dari pemodelan topik hasil penelitian yaitu terbagi menjadi dua topik yaitu topik umum dan topik penyakit. Hasil pengujian penelitian ini 94,1% mengatakan sangat baik [9]. Penelitian terkait

LDA oleh [5] pada data abstrak skripsi dipersiapkan melalui tahap *preprocessing* untuk mempermudah dalam topic modelling. Hasil dari preprocessing kemudian dihitung jumlah kemunculan setiap kata dengan model *bag of words*. Gabungan metode NBC dan LDA untuk mengetahui sentimen pengguna aplikasi Ruang Guru untuk mengetahui topik yang dibicarakan pengguna aplikasi tersebut sebagai bahan evaluasi perusahaan[10]. Penelitian ini bertujuan melihat hasil tingkat akurasi pemodelan topik LDA yang diberikan sentimen analisis menggunakan NBC,

2 Metodologi

Penelitian ini terdapat beberapa proses yang harus dilewati sebagai alur penelitian seperti yang dilukan dalam beragai penelitian serupa. Secara umum langkah penelitian ini dilakukan seperti yang tertampil pada Gambar 1



Gambar 1 Alur metode penelitian.

3 Hasil dan pembahasan

Pada penelitian ini telah di tambang sebanyak 5000 data tweet, data tweet yang diunduh autentik dengan data tweet yang diunggah oleh pengguna twitter. Langkah *preprocessing* dilakukan dengan proses *lower case*, menghapus teks yang berupa url atau link, menghapus tanda baca seperti karakter !#$%&'()*+,- ./:;<=>?@[] dan menghilangkan kata yang tidak perlu digunakan agar mempermudah proses klasifikasi *stopword*.

NBC merupakan teknik pembelajaran algoritma data mining yang memanfaatkan metode probabilitas dan statistik. NBC dalam melakukan klasifikasi terdapat dua proses penting yaitu *learning* (*training*) dan *testing* [11]. Penelitian ini menggunakan data training dari *tweets* yang diperoleh yaitu sebanyak 1500 data *tweet* dengan kategori opini positif 500 *tweet*,

opini negatif 500 tweet dan opini netral 500 tweet beserta klasifikasinya secara manual atau data *tweet* yang sudah diketahui kategorinya.

Proses klasifikasi NBC dengan mempresentasikan setiap *tweet* dengan atribut $x_1, x_2, \ldots, x_n$ dengan x_1 untuk kata pertama, x_2 adalah kata kedua, dan seterusnya. Untuk himpunan kategori *tweet* dipresentasikan dengan V. Saat melakukan proses klasifikasi, NBC akan mencari nilai probabilitas tertinggi dari semua kategori dokumen yang diujikan (V_{MAP}) yang dihitung menggunakan Persamaan 1.

$$V_{MAP} = \arg\max_{v_j \in V} = \frac{P(x_1, x_2, \ldots, x_n|v_j)(Pv_j)}{P(x_1, x_2, \ldots, x_n|v_j)} \quad (1)$$

Karena nilai $P(x_1, x_2, ..., x_n)$ adalah konstan untuk semua kategori v_j, sehingga dapat ditulis menjadi Persamaan 2.

$$V_{MAP} = \arg\max_{v_j \in V} = P(x_1, x_2, \ldots, x_n|v_j)P(v_j) \quad (2)$$

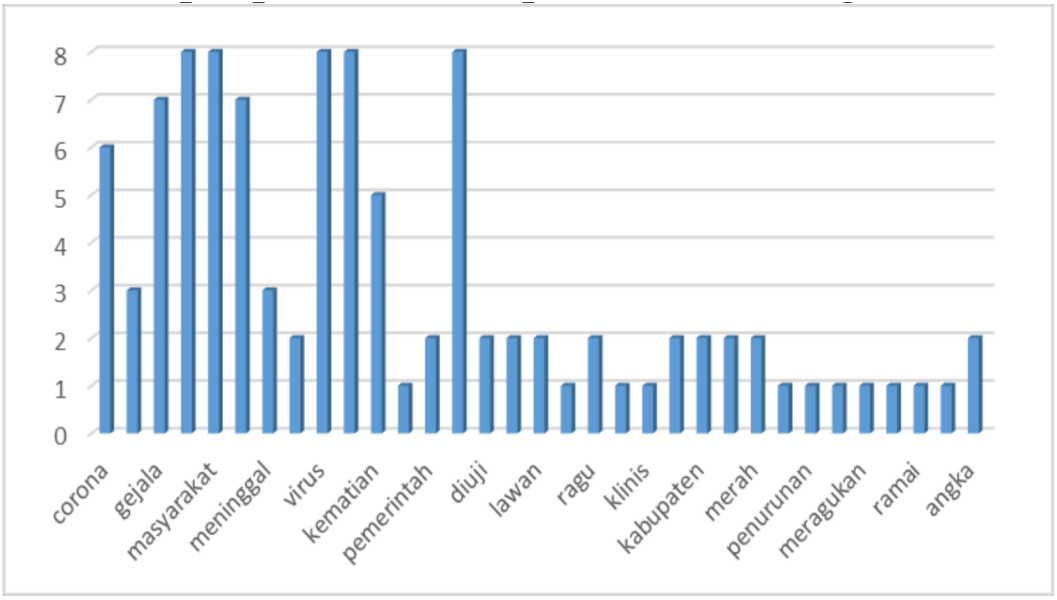
LDA merupakan algoritma sederhana untuk topic modeling [6], dalam *machine learning*, LDA termasuk dalam kelompok *unsupervised learning*. LDA muncul sebagai salah satu metode yang digunakan untuk melakukan analisis pada dokumen yang sangat besar. Untuk setiap dokumen dalam koleksi, algoritma LDA mengambil topik berdasarkan distribusi kata-kata multinomial dalam dokumen. Kemudian topik tersebut digunakan untuk menghasilkan kata itu sendiri berdasarkan distribusi multinomial topik dan mengulangi dua langkah ini untuk semua kata dalam dokumen [5]. Pada sampel sentimen yang diklasifikasi dengan NBC, sebuah tweet yang belum diketahui sentimennya, yakni kalimat "vaksin berhasil sembuh". Perbandingan nilai probabilitas kata "vaksin" , "berhasil", dan "sembuh" pada setiap sentimennya memiliki nilai positif dengan nilai 0.000528 dan nilai negatif sebesar 0.00000006.

LDA yang merupakan *unsupervised learning*, pada penelitian ini hasil sebaran maupun frekuensi hasil pemodelan topik LDA dari manusia sebagai Interpreter menghasilkan contoh frekuensi kata pada topik negatif tertampil dalam Gambar 2. Contoh dari pemodelan topik menggunakan LDA adalah seperti tertampil dalam Tabel1. Hasil dari topik positif pada pemodelan LDA dengan topik Kesehatan, menghasilkan sebuah topik kesehatan yang mempunyai tanggapan positif dari masyarakat pengguna twiter berkaitan dengan dukungan terhadap penanganan Covid-19. Sentimen positif juga berkaitan dengan program vaksin yang diselenggarakan Pemerintah.

Tabel 1 Pemodelan topik LDA.

No	Positif	Netral	Negatif
1	Presiden Program Vaksin	Vaksin di Kota dan Kabupaten Aceh	Masyarakat Meragukan Vaksin Virus Corona
2	Pertamina Peduli Program 5M	Agenda Pemerintah Pusat	Angka Kematian Kota Zona Merah
3	Mendonorkan Berbagi Antibodi	Penularan Virus Corona	Keamanan Vaksin Diuji Pemerintah

Pada bulan April 2021 topik kesehatan yang memiliki sentimen positif berkaitan dengan masyarakat yang mendukung untuk meminimalisir peningkatan Covid-19 dengan menerapkan libur lebaran yang dilakukan dirumah. Pada bulan yang sama topik kesehatan yang memiliki sentimen netral berkaitan dengan penurunan rantaian penularan covid-19. Sedangkan topik kesehatan yang memiliki sentimen negatif berkaitan dengan efek negatif pelanggaran protocol



Gambar 2 Frekuensi kata pada topik negatif .

kesehatan pada masa mudik lebaran. Hasil sentimen NBC yang memiliki akurasi paling baik diujikan sebagai dataset LDA dihasilkan pada iterasi ke-7. Hasil akurasi yang didapatkan adalah 70% dengan tingkat presisi 90.8% dan *recall* 89.09%. Hasil data sentimen tersebut digunakan sebagai dataset pada pemodelan topik menggunakan LDA yang telah didapatkan tingkat akurasi 78% dengan 10 topik. Hasil topik kesehatan yang diunggah pengguna twitter dapat dianalisis secara spesifik sesuai dengan sentimen dengan NBC dapat diidentifikasi topiknya dengan nilai akurasi akhir sebasar 83.3%.

4 Kesimpulan

Penelitian ini memiliki akurasi yang tinggi dalam klasifikasi sentimen twitter yang berkaitan dengan kesehatan menggunakan NBC dengan hasil akurasi 99,7 %. Pemodelan topik data tweet tentang kesehatan dengan menggunakan LDA dihasilkan topik lebih spesifik yang memiliki akurasi 78%. Dengan dilakukannya klasifikasi sentimen menggunakan NBC sebelum pemodelan topik dengan LDA. Analisis secara spesifik sesuai dengan sentimen dan dapat diidentifikasi topiknya dengan memiliki nilai akurasi akhir 83.3% atau terjadi peningkatan sebesar 5.3%.

Pustaka

1 C. Sohrabi, Z. Alsafi, N. O'neill, M. Khan, A. Kerwan, A. Al-Jabir, C. Iosifidis, dan R. Agha, "World health organization declares global emergency: A review of the 2019 novel coronavirus (covid-19)," *International journal of surgery*, vol. 76, pp. 71–76, 2020.

2 P. Antinasari, R. S. Perdana, dan M. A. Fauzi, "Analisis sentimen tentang opini film pada dokumen twitter berbahasa indonesia menggunakan naive bayes dengan perbaikan kata tidak baku," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN*, vol. 2548, p. 964X, 2017.

3 R. Y. Hayuningtyas dan R. Sari, "Analisis sentimen opini publik bahasa indonesia terhadap wisata tmii menggunakan naïve bayes dan pso," *Jurnal Techno Nusa Mandiri*, vol. 16, no. 1, pp. 37–42, 2019.

4 H. Schütze, C. D. Manning, dan P. Raghavan, *Introduction to information retrieval*. Cambridge University Press Cambridge, 2008, vol. 39.

5 A. I. Alfanzar, K. Khalid, dan I. S. Rozas, "Topic modelling skripsi menggunakan metode latent diriclhet allocation," *JSiI (Jurnal Sistem Informasi)*, vol. 7, no. 1, pp. 7–13, 2020.

6 K. Arfianti, "Identifikasi topik artikel berita menggunakan topic modelling dengan metode latent dirichlet allocation," 2019.

7 A. Sudiantoro dan E. Zuliarso, "Analisis sentimen twitter menggunakan text mining dengan algoritma naive bayes classifier," *Jurnal Dinamika Informatika*, vol. 10, no. 2, pp. 69–73, Oct. 2018. [Online]. Available: https://unisbank.ac.id/ojs/index.php/fti2/article/view/8135

8 Y. Mahardhika dan E. Zuliarso, "Analisis sentimen terhadap pemerintahan joko widodo pada media sosial twitter menggunakan algoritma naive bayes classifier," *SINTAK*, vol. 2, Nov. 2018. [Online]. Available: https://unisbank.ac.id/ojs/index.php/sintak/article/view/6651

9 Y. Sahria, D. H. Fudholi *et al.*, "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, 2020.

10 M. R. Firdaus, F. M. Rizki, F. M. Gaus, dan I. K. Susanto, "Analisis sentimen dan topic modelling dalam aplikasi ruangguru," *J-SAKTI (Jurnal Sains Komputer dan Informatika)*, vol. 4, no. 1, pp. 66–76, 2020.

11 F. Nurhuda, S. W. Sihwi, dan A. Doewes, "Analisis sentimen masyarakat terhadap calon presiden indonesia 2014 berdasarkan opini dari twitter menggunakan metode naive bayes classifier," *ITSmart: Jurnal Teknologi dan Informasi*, vol. 2, no. 2, pp. 35–42, 2013.