

PENERAPAN METODE *TOPIK MODELING* UNTUK PENENTUAN TOPIK KONSULTASI PADA PORTAL TELEMEDICINE MENGGUNAKAN LDA (*LATENT DIRCHLECT ALLOCATION*)

Saeful Bahri¹, Gunawan², Dede Wintana³, Rusda Wajhillah⁴, Satia Suhada⁵

Universitas Bina Sarana Informatika

Jalan Cemerlang No. 8 Sukakarya Kota Sukabumi, telp/fax 6251992, 6251993

saeful.sel@bsi.ac.id¹, gunawan.gnz@bsi.ac.id², Dede.dwe@bsi.ac.id³

rusda.rwh@bsi.ac.id⁴, Satia.shq@bsi.ac.id

Abstract

The Novel Corona Virus or Covid-19 pandemic that occurred at the beginning of 2020 has had a very big change for the medical world around the world, changes occurred in the pattern of hospital medical services and the work patterns of the medical personnel themselves, this was due to the fear of being infected with the virus. , one of the alternatives provided by the medical world is the application of the telemedicine method, but in practice telemedicine has many shortcomings, one of which is the difficulty of determining the topics in health discussed because most telemedicine documents are usually in text format, this makes it a challenge for researchers in processing text to find something. which can be used to improve consultation outcomes in a telemedicine system. One approach that is quite popular and powerful in finding themes in the medical and health corpus is topic modeling, one of which uses LDA (Latent Dirichlect Allocation), by applying LDA to the determination of TOPIK in the case of corpus telemedicine, it is proven to be able to show good results in terms of value. the highest kohence is 0.551075 with the highest standard deviation with a value of 0.5327286.

Keywords: Covid-19; Telemedicine; Pandemic; LDA;

Abstrak

Pandemi Novel Corona Virus atau covid-19 yang terjadi pada awala tahun 2020 telah berdampak perubahan yang sangat besar bagi dunia medis diseluruh dunia, perubahan terjadi pada pola pelayanan medis rumah sakit dan pola kerja tenaga medis itu sendiri, hal ini disebabkan karena ketakutan akan terinfeksi virus, salah satu alternatif yang diberikan oleh dunia medis adalah dengan penerapan metode telemedicine namun dalam prakteknya telemedicine memiliki banyak kekurangan salahsatunya seperti sulitnya penentuan topik dalam kesehatan yang dibahas karena kebanyakan dokument telemedicine biasanya berformat text, hal tersebut menjadikan tantangan bagi peneliti dalam memproses text dalam menemukan sesuatu yang bermakna sehingga dapat dimanfaatkan untuk meningkatkan hasil konsultasi dalam sebuah sistem telemedicine. Salah satu pendekatan yang cukup populer dan powerful dalam penemuan tema dalam korpus medis dan kesehatan adalah pemodelan topik, salah satunya menggunakan LDA (Latent Dirichlect Allocation),

dengan diterapkan nya LDA pada penentuan TOPIK dalam kasus korpus telemedicine terbukti mampu menunjukan hasil yang baik dilihat dari nilai kohence tertinggi yaitu 0,551075 dengan standar deviasi tertinggi dengan nilai 0,5327286.

Kata kunci: Covid-19; Telemedicine; Pandemic; LDA;

1. PENDAHULUAN

Di Awal tahun 2020 dunia dikejutkan dengan mewabahnya virus corona yang pertama kali ditemukan Desember 2019, di kota Wuhan, provinsi Hubei, China [1]. Kemudian pada bulan April 2020 WHO menetapkan corona sebagai pandemic dunia, pandemic Novel Corona Virus atau covid-19 memiliki dampak perubahan yang sangat besar bagi dunia medis diseluruh dunia, perubahan terjadi pada pola pelayanan medis rumah sakit dan pola kerja tenaga medis itu sendiri, hal ini disebabkan karena ketakutan akan terinfeksi virus[2], salah satu alternatif yang ditawarkan oleh dunia medis adalah dengan penerapan metode telemedicine atau pengobatan jarak jauh, dengan menggunakan beberapa media seperti telepon, internet dan sistem yang terintegrasi dengan rumah sakit itu sendiri[3][4], metode telemedicine terbukti mampu mengurangi resiko dalam penyebaran wabah seperti covid-19 [5][6].

Banyaknya data dalam format digital akan sangat menyulitkan manusia untuk mencari dan menemukan pola dalam data khususnya data medis dan kesehatan sebagai contoh peningkatan 600% pada website layanan kesehatan seperti halodoc dan alodoc[1]. Hal tersebut menjadi motivasi peneliti dalam menemukan pola dan informasi dari data tersebut.

Perkembangan teknik yang efisien dalam melakukan analisa dalam menemukan pola dan informasi dari dataset medis yang cukup besar telah menjadi trend saat diantanya menggunakan. Teknologi kecerdasan buatan, teknologi ini telah berkembang pesat dan telah banyak diterapkan dalam dunia medis khususnya pada metode telemedicine hal ini terlihat dari beberapa penelitian yang telah dilakukan diantaranya penelitian yang dilakukan oleh[12] yang melakukan penelitian tentang *framework* kecerdasan buatan yang digunakan dalam melakukan klasifikasi suara paru-paru pada penderita gangguan paru-paru, dalam penelitiannya dikemukakan bahwa kecerdasan buatan dapat dengan cepat menilai kondisi paru-paru seorang pasien, kemudian dalam penelitian yang dilakukan oleh[4] dalam penelitiannya menyatakan bahwa kecerdasan buatan dapat digunakan untuk melakukan diagnostic pada pasien dengan gangguan bicara. Sementara itu dalam penelitian yang dilakukan oleh [13] dalam penelitiannya mengemukakan bahwa penerapan kecerdasan buatan merupakan salah satu cara untuk memaksimalkan metode telemedicine, karena dengan teknologi kecerdasan buatan dokter atau pasien dapat dengan mudah memaintenance status Kesehatan dengan relatif lebih mudah.

Salah satu metode yang populer dalam pemodelan topik yaitu Latent Dirichlet Allocation (LDA) metode digunakan untuk mendeteksi topik-topik yang ada pada koleksi dokumen beserta proporsi kemunculan topik tersebut, baik di koleksi maupun di dokumen tertentu, LDA mampu mengasosiasikan kata-kata yang ada.

2. METODOLOGI PENELITIAN

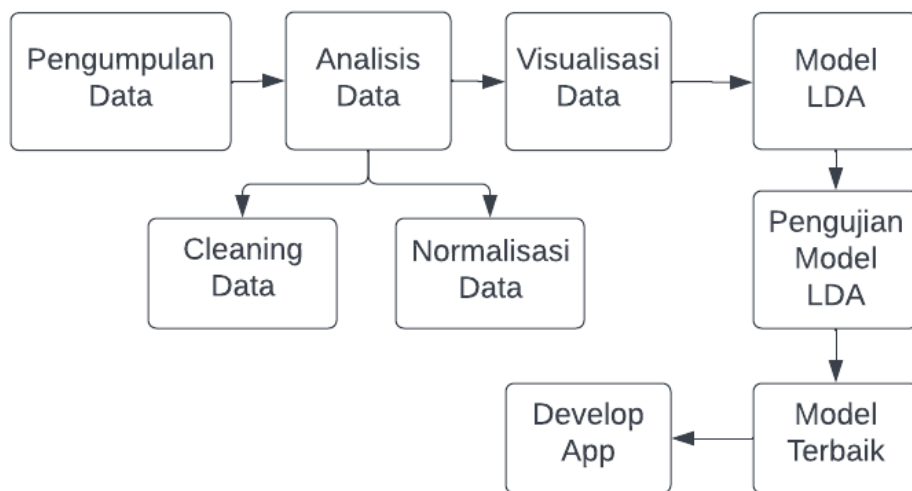
2.1. Penelitian Terkait

Text minning dapat didefinisikan sebagai sebuah metode dalam machine learning dan statistik dengan tujuan pengenalan pola dan penemuan informasi yang tersembunyi dalam sekumpulan text atau corpus. Pada bagian ini peneliti akan meninjau konsep-konsep penting dan penerapan topik modeling menggunakan LDA (Latent Dirhlect Alocation) . Terdapat dua penelekatan utama dalam data minning yaitu supervised learning dan unsupervised learning. Supervised learning memiliki tujuan menemukan pola dari dataset yang memiliki label sedangkan pada Unsupervised learning pada penemuan polanya menggukan dataset yang tidak memiliki label. Salah satu teknik yang populer dalam pendekatan dengan supervised learning dan unsupervised learning adalah klasifikasi dan klustering. Sesuai dengan pengertian tersebut klasifikasi dapat digunakan untuk melakukan training pada korpus yang telah terdapat label, sedangkan klastering berguna untuk menetapkan klaster untuk setiap dokument berdasarkan kesamaan dan perbedaan antar klaster. Diantara teknik dalam text minning pemodelan topik adalah salah satu model yang populer digunakan untuk metode unsupervised learning, beberapa penggunaan yang populer adalah Deteksi spam dan image tagging. LDA merupakan sebuah model Bayesian hirarki tingkat tiga . LDA diassumsikan berisi latent topik dan setiap topik dapat diwakili oleh distribusi diseluruh kata-kata[6].

LDA telah diterapkan dalam bidang medis dan kesehatan diantaranya memprediksi hubungan protein-protein berdasarkan ilmu sastra[4] menemukan konsep dan struktur klinis yang relevan dalam catatan kesehatan pasien [3], mengidentifikasi pola kejadian klinis dalam kohort pasien kanker otak [2], dan menganalisis hasil waktu-ke-kejadian [11]. Penemuan pola Clinical Pathway (CP) adalah metode untuk mengungkapkan struktur, semantik, dan dinamika CP untuk memberikan dokter dengan pengetahuan eksplisit yang digunakan untuk memandu kegiatan pengobatan pasien individu. LDA telah digunakan untuk CP untuk menemukan perilaku pengobatan pasien [14], untuk memprediksi pola urutan klinis, dan untuk model berbagai kegiatan pengobatan [7] dan cap waktu yang terjadi di CP [12]. LDA juga telah disesuaikan untuk menentukan kematian pasien [15], dan untuk menemukan pengetahuan dari pemodelan penyakit dan karakteristik pasien [16]. Redundancyaware LDA (Red-LDA) adalah salah satu versi LDA untuk menangani masalah redundansi dalam dokumen medis dan telah menunjukkan kinerja yang lebih baik daripada LDA [9].

2.2. Alur Penelitian

Selain penelitian terdahulu pada penelitian ini juga kami merancang tahapan dari peneltian tahapan yang digunakan dalam penelitian ini menggunakan metode penelitian kuantitatif, yaitu akan dijelaskan dalam gambar 1 berikut



Gambar. 1 Alur Penelitian

2.2.1. Pengumpulan Data

Metode pengumpulan data yang dilakukan dalam penelitian ini adalah dengan mengambil data berupa corpus text dari beberapa web layanan kesehatan seperti halodoc.com dan berapa website layanan kesehatan yang beredar dan telah berlisensi dari kementerian keshatan dan kementerian informatika republik indonesia, korpus yang didapat dialih formatkan kedalam format .txt untuk selanjutnya data tersebut dianalisa pada tahapan berikutnya.

2.2.2. Analisa Data

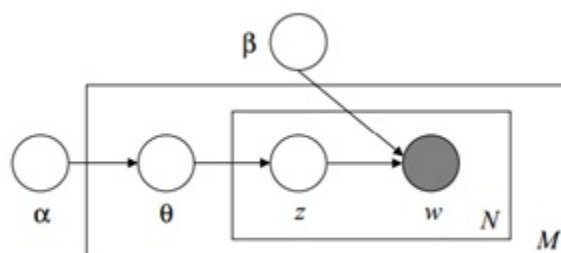
Corpus yang telah terkumpul selajutnya dilakukan proses analisis untuk membersihkan data dari beberapa tanda baca, simbol, karakter khusus, duplikasi dan menyeragamkan format text yang ada dalam korpus tersebut. Setelah corpus dibersihkan selanjutnya korpus dinormalisi dengan cara melakukan pemrosesan text seperti proses yang meliputi diantaranya proses tokenisasi, *stopword removal*, *stemming*. Dalam proses tokenisasi semua text yang terdapat dalam corpus akan dirubah bentuk kedalam *all lower case*, setelah proses tokenisasi selanjutnya dilakukan proses penghapusan kata-kata yang kurang bermakna, biasanya proses ini akan dilakukan dengan berpatokan pada kamus kata dasar, selanjutnya corpus yang telah diproses dengan tokenisasi dan *stopword removal* akan diproses kembali untuk mencari root kata dengan proses stemming.

2.2.3. Visualisasi Data

Visualisasi data diperlukan untuk menggambarkan bagai mana data tersebut, visualisasi dalam penelitian ini digunakan untuk mempermudah dalam mengalisa tren, outlier dan pemahaman pola pada corpus yang digunakan sehingga dapat diminimalisir kesalahan dalam proses analisa dan kegagalan dalam pembentukan model.

2.2.4. Pembuatan Model

Dalam pembuatan model pada penelitian ini menggunakan model LDA atau Latent Dirichlet Allocation yang merupakan sebuah model probabilistik generatif, LDA sendiri merupakan sebuah model hirarki bayesian yang memiliki tiga tingkat dimana setiap koleksi item dimodelkan sebagai campuran terbatas atas serangkaian set topik.



Sumber: <https://socs.binus.ac.id/files/2018/11/dw-1.jpg>

Gambar. 2. Represensi Model LDA

2.2.5. Penentuan Jumlah Topik

Sebelum dilakukann pengujian model kami melakukan experiment untuk menentukan model topik terbaik, tahapan ini menjadi sangat penting karena untuk mendapatkan nilai koherence yang tinggi, karena semakin tinggi nilai koherence maka akurasi model semakin baik. Penentuan jumlah topik dilakukan dengan cara melakukan modifikasi pada jumlah topik, pemberian jumlah topik dalam penelitian ini dilakukan sebanyak 3 kali yaitu 2,3 dan 4 topik, setelah terlihat tren nilainya maka akan dilakukan perhitungan standar deviasinya. Kemudian dipilih jumlah topik dengan nilai rata-rata standar deviasi paling tinggi perbandingan nilai Koherence dapat diliaht pada tabel 1.

Tabel 1. Experiment Penentuan Jumlah Topik

2 Topik		3 Topik		4 Topik	
Topik	Nilai Koherence	Topik	Nilai Koherence	Topik	Nilai Koherence
1	0.606909	1	0.569430	1	0.53058
2	0.482873	2	0.589417	2	0.53562
		3	0.494378	3	0.56248
				4	0.51672

Berdasarkan experiment dari 3 parameter topik yang digunakan didapatkan nilai koheren masing masing pada pengulangan 2 topik nilai koherence tertinggi pada topik ke 1 dengan nilai 0.606909, pada pengulangan 3 topik nilai koheence tertinggi adalah pada topik ke 2 dengan nilai 0.589417 sedangkan pada pengulangan dengan 3 topik nilai terbesar terdapat pada topik ke 3 dengan nilai koherence sebesar

0.56248. nilai koherence yang didapat dari masing-masing pengulangan selanjutnya di cari rata-rata sesuai dengan topik pada masing-masing pengulangan. Ilai tersebut digunakan untuk mencari topik dari penyakit yang paling banyak dikonsultasikan oleh masyarakat melalui platform layanan kesehatan online. Berdasrkan perhitungan rata-rata didapat topik ke 2 memiliki rata nilai kohence tertinggi yaitu 0,551075 dengan standar deviasi tertinggi dengan nilai 0,5327286.

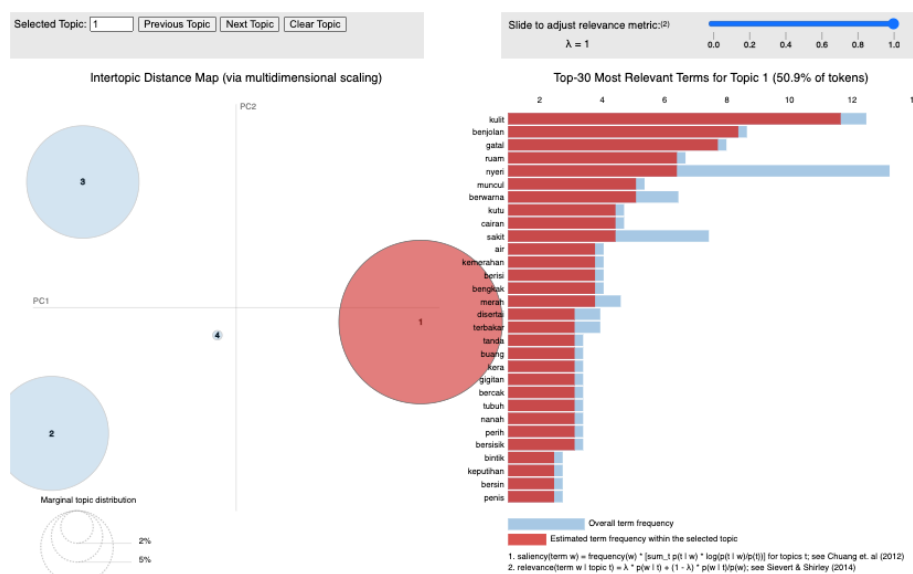
3. HASIL DAN PEMBAHASAN

Setelah mendapatkan jumlah topik berdasarkan percobaab sebelumnya maka topik yang didpat adalah sebanyak 4 topik dengan masing-masing nilai kata memiliki bobot tersediri, hasil dapt dilihat pada table 2 hasil pemodelan topik.

Tabel 2. Hasil Pemodelan Topik

Topik	Pemodelan Topik
Topik 1	'0.004*"napa" + 0.004*"panas" + 0.004*"pecahpecah" + 0.004*"area"'
Topik 2	'0.035*"nyeri" + 0.029*"napas" + 0.023*"batuk" + 0.023*"dada"'
Topik 3	'0.004*"kulit" + 0.004*"napas" + 0.004*"sesak" + 0.004*"nyeri"'
Topik 4	'0.044*"kulit" + 0.031*"benjolan" + 0.028*"gatal" + 0.022*"ruam"'

Berdasarkan data tersebut dapat divisualisasikan kembali kedalam pyLDvavis untuk melihat keterkaitan antar kata yang dihasilkan sehingga keterhubungan antar kata dapat digunakan dalam pemodelan topik

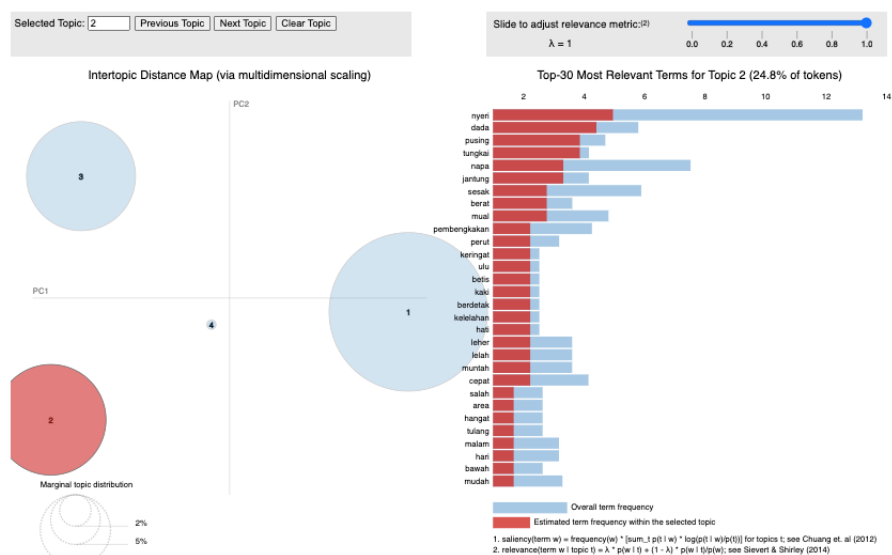


Gambar. 3. Visualisasi pyDALvis Topik Ke 1

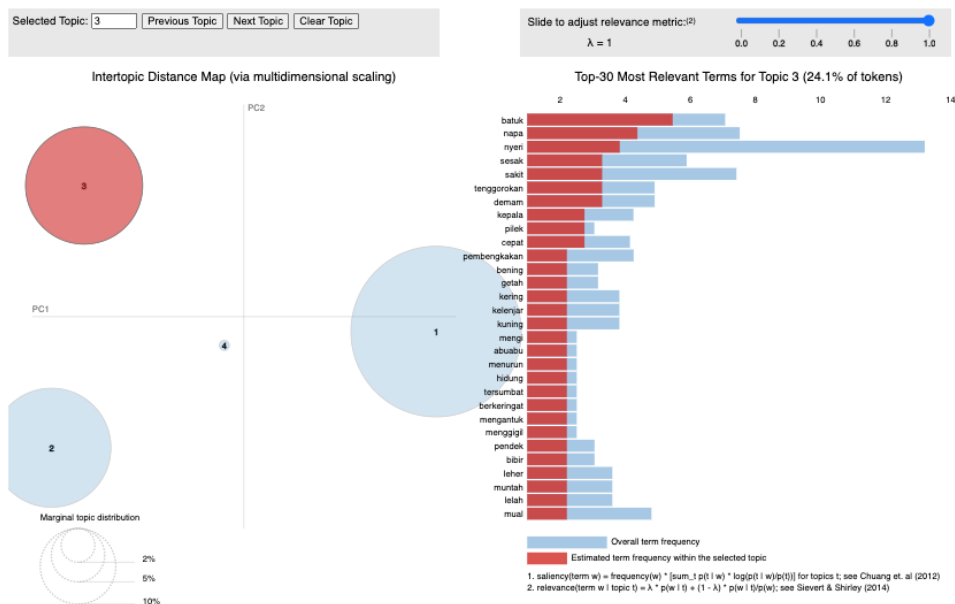
Berdasarkan gambar 3 tentang visualisasi pyDALvis.

Topik ke 1 didapatkan kata-kata yang sering muncul dan berkaitan satu dengan yang lain adalah kata-kata kulit, benjolan, gatal, ruam, nyeri, muncul, kutu cairan, sakit, bengkak, merah, gigitan, nanah, bercak, bersisik sehingga dari kata-kata tersebut dapat direpresentasikan menjadi pembahasan sebagai "Sakit kulit

menyebabkan gagal-gatal”, pembahasan yang paling dominan dalam topik ke 1 adalah pembahasan tentang penyakit yang menyerang kulit.



Gambar. 4 Visualisasi pyDALvis Topik Ke 2



Gambar. 5. Visualisasi pyDALvis Topik ke 3

Topik ke 2 dan ke 3 didapatkan kata-kata yang sering muncul dan berkaitan satu dengan yang lain adalah kata-kata seperti nyeri, dada, sesak nafas, batuk berdehak, kelelahan, berkerengat malam hari, sehingga dari kata-kata tersebut dapat direpresentasikan menjadi pembahasan sebagai “Nyeri dada batuk berdehak sesak nafas”, dapat kalimat tersebut , adapun kata-kata pada topik ke 3 kata yang sering

muncul relatif sama dengan topik ke 2 sehingga topik ke 3 dapat di asosiasikan kedalam pembahasa tentang penyakit pernapasan. Sedangkan untuk topik ke 4 tidak memiliki keterhubungan antar kata sehingga tidak dapat digunakan untuk pemodelan topik

4. SIMPULAN

Berdasarkan hasil pembahasan diatas maka dapat diambil beberapa kesimpulan diantaranya:

- a. Pemodelan topik menggunakan LDA (Latent Dirhlect Allocation) menggunakan data/corpus dari portal layanan kesehatan, corpus berupa pertanyaan dan jawaban seputar kesehatan masyarakat, selanjutnya corpus dilakukan praprocessing untuk membersihkan corpus dari kata yang kurang bermakna, selajutnya dilakukan percobaan untuk menetetukan topik terbaik dengan metode pengukuran *score coherence*, dari percobaan tersebut didapat skor coheren terbaik pada topik ke 2 dengan nilai rata sebesar 0,551075 dengan standar deviasi tertinggi sebesar 0,5327286.
- b. Hasil dari analisis pemodelan menggunakan metode LDA (Laent Dirichlect Allocation) menghasilkan 2 kesimpulan diantaranya :
- c. Pada topic ke 1 topik yang sering ditanyakan masyrakat adalah tentang penyakit yang menyerang kulit, seperti gatal-gatal, benjolan dan ruam pada kulit
- d. Pada topic ke 2 dan topik ke 3 topik yang sering ditanyakan oleh masyarakat yaitu tentang penyakit yang menyerang pernapasan, seperti sesak dan batuk

DAFTAR PUSTAKA

- [1] I. Chakraborty and P. Maity, "COVID-19 outbreak: Migration, effects on society, global environment and prevention," *Sci. Total Environ.*, vol. 728, p. 138882, 2020, doi: 10.1016/j.scitotenv.2020.138882.
- [2] W. H. Chang, "The influences of the COVID-19 pandemic on medical service behaviors," *Taiwan. J. Obstet. Gynecol.*, vol. 59, no. 6, pp. 821–827, 2020, doi: 10.1016/j.tjog.2020.09.007.
- [3] W. R. Smith, A. J. Atala, R. P. Terlecki, E. E. Kelly, and C. A. Matthews, "Implementation Guide for Rapid Integration of an Outpatient Telemedicine Program During the COVID-19 Pandemic," *J. Am. Coll. Surg.*, vol. 231, no. 2, pp. 216-222.e2, 2020, doi: 10.1016/j.jamcollsurg.2020.04.030.
- [4] J. R. Duffy, G. W. Werven, and A. E. Aronson, "Telemedicine and the diagnosis of speech and language disorders," *Mayo Clin. Proc.*, vol. 72, no. 12, pp. 1116–1122, 1997, doi: 10.4065/72.12.1116.
- [5] A. S. Albahri *et al.*, "IoT-based telemedicine for disease prevention and health promotion: State-of-the-Art," *J. Netw. Comput. Appl.*, vol. 173, p. 102873, 2021, doi: 10.1016/j.jnca.2020.102873.

- [6] M. Abuzeineh *et al.*, "Telemedicine in the Care of Kidney Transplant Recipients With Coronavirus Disease 2019: Case Reports," *Transplant. Proc.*, vol. 52, no. 9, pp. 2620–2625, 2020, doi: 10.1016/j.transproceed.2020.07.009.
- [7] L. Ting, B. M. Bs, and M. Wilkes, "Telemedicine for Patient Management on Expeditions in Remote and Austere Environments : A Systematic Review," *Wilderness Environ. Med.*, no. February 2018, 2020, doi: 10.1016/j.wem.2020.09.009.
- [8] A. Ataç, E. Kurt, and S. E. Yurdakul, "An Overview to Ethical Problems in Telemedicine Technology," *Procedia - Soc. Behav. Sci.*, vol. 103, pp. 116–121, 2013, doi: 10.1016/j.sbspro.2013.10.315.
- [9] B. Weiss, N. Paul, F. Balzer, D. T. Noritomi, and C. D. Spies, "Telemedicine in the intensive care unit: A vehicle to improve quality of care?," *J. Crit. Care*, vol. 61, pp. 241–246, 2021, doi: 10.1016/j.jcrc.2020.09.036.
- [10] F. Hentati, C. I. Cabrera, and B. D. Anza, "Patient satisfaction with telemedicine in rhinology during the COVID-19 pandemic," *Am. J. Otolaryngol. Neck Med. Surg.*, p. 102921, 2021, doi: 10.1016/j.amjoto.2021.102921.
- [11] J. P. O. Li *et al.*, "Digital technology, tele-medicine and artificial intelligence in ophthalmology: A global perspective," *Prog. Retin. Eye Res.*, p. 100900, 2020, doi: 10.1016/j.preteyeres.2020.100900.
- [12] M. M. Jaber, S. K. Abd, P. M. Shakeel, M. A. Burhanuddin, M. A. Mohammed, and S. Yussof, "A telemedicine tool framework for lung sounds classification using ensemble classifier algorithms," *Meas. J. Int. Meas. Confed.*, vol. 162, p. 107883, 2020, doi: 10.1016/j.measurement.2020.107883.
- [13] B. F. Oladejo, "Improving Medical Consultation Process through Natural Language Processing Technique," vol. 11, no. 7, pp. 1760–1766, 2020.
- [14] J. Coquet *et al.*, "Comparison of orthogonal NLP methods for clinical phenotyping and assessment of bone scan utilization among prostate cancer patients," *J. Biomed. Inform.*, vol. 94, no. April, p. 103184, 2019, doi: 10.1016/j.jbi.2019.103184.
- [15] E. Sarioglu, H.-A. Choi, and K. Yadav. Clinical report classification using natural language processing and topic modeling. In *Proceedings of the IEEE International Conference on Machine Learning and Applications (ICMLA)*, volume 2, pages 204–209, 2012.
- [16] C. Arnold and W. Speier. A topic model of clinical reports. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1031–1032. ACM, 2012.