Deteksi Komunitas Berdasarkan Kemiripan *Tweet* pada Penerapan Protokol Kesehatan di DKI Jakarta Menggunakan Algoritma Louvain

Bimo Satrio Aji_1301171248, Ilham Wahyu Adli_1301173380

Pada tahun 2020, Indonesia menjadi salah satu negara dengan kasus COVID-19 yang tinggi. Pada Desember 2020, angka kematian akibat kasus COVID-19 di Indonesia telah mencapai lebih dari 3% dari jumlah kasus positif. Jumlah kasus ini semakin hari semakin meningkat, karena kurangnya antisipasi yang dilakukan pemerintah dan masyarakat. Saat ini, pemerintah dengan usahanya sudah banyak melakukan banyak kegiatan untuk terus memberikan pengetahuan dan pencegahan untuk mengatasi COVID-19. Salah satu usaha tersebut adalah dengan cara melakukan kampanye tentang protokol kesehatan, disisi lain banyak aktor yang ikut terlibat dalam kampanye tersebut, seperti pengguna jejaring sosial media Twitter yang memiliki banyak *followers* sehingga dapat memberikan banyak pengaruh bagi orang-orang yang mengikutinya. Tujuan dari penelitian ini adalah untuk medeteksi komunitas yang muncul dari adanya kampanye tentang protokol kesehatan yang dilakukan di media sosial Twitter. Komunitas itu terjadi karena adanya kemiripan *tweet* yang dibuat oleh *user*, kemiripan ini dapat terlihat terkait dengan konteks dari *keyword* yang dipilih dalam pengambilan dataset. Penelitian ini diharapkan dapat membantu masyarakat khususnya pengguna Twitter untuk mengetahui tentang isu yang sedang berkembang terkai sosialisasi protokol kesehatan. Pada penelitian ini data yang digunakan adalah dataset dari jejaring sosial media Twitter, dengan kata kunci terkait protokol kesehatan yang dikampanyekan pemerintah. Data hasil *scraping* ini nantinya akan dilihat kemiripannya menggunakan algoritma Cosine Similarity dan mendeteksi komunitasnya menggunakan algoritma Louvain. Penelitian ini menghasilkan komunitas dengan jumlah node paling banyak sebesar 44 yang memiliki bahasan tentang cuci tangan.

Kata Kunci—Twitter, COVID-19, Virus, Akun, Protokol, Kesehatan, Deteksi Komunitas

I. PENDAHULAN

Coronavirus (CoV) adalah keluarga besar virus yang menyebabkan penyakit mulai dari flu biasa hingga penyakit yang lebih parah seperti Middle East Respiratory Syndrome (MERS-CoV) dan Severe Acute Respiratory Syndrome (SARS-CoV). Penyakit Coronavirus (COVID-19) adalah jenis baru yang ditemukan pada bulan Desember 2019 di Wuhan, Cina dan belum pernah diidentifikasi pada manusia [1]. World Health Organization (WHO) dalam situs resminya menyatakan bahwa COVID-19 dikategorikan sebagai sebuah pandemik [2].

Menurut WHO mempraktikan kebersihan diri sangat penting seperti cuci tangan merupakan cara terbaik untuk melindungi diri sendiri dan orang lain. Jika memungkinkan, melakukan jaga jarak setidaknya satu meter antara diri sendiri orang lain juga merupakan sangat penting [3]. Kemenkes juga telah menerbitkan protokol kesehatan di tempat umum. Penggunaan masker awalnya hanya disarankan ketika seseorang merasa sakit, tetapi hal ini dirubah pada peraturan tersebut untuk menekan angka pertumbuhan penyebaran COVID-19 [4]. Kampanye tentang protokol kesehatan ini juga dilakukan oleh banyak pihak termasuk para pengguna sosial media Twitter. Dibalik banyaknya kampanye dan informasi yang tersebar ini ada komunitas yang membahas suatu topik tertentu dalam *tweet*-nya. Komunitas tersebut menjadi informasi penting apa saja topik yang paling bahas dibahas pada masa pandemi COVID-19.

II. PENELITIAN TERKAIT

Akhmad Irsyad dan Nur Aini Rakhmawati menggunakan algoritma Louvain untuk mendeteksi komunitas di media sosial Twitter dengan melihat kesamaan tweet terkait yang menghasilkan kelompok yang akhirnya menjadi suatu komunitas [5]. Pada studi lainnya yang telah dilakukan oleh Abdelsadek et al., 2018 yaitu mengidentifikasi komunitas berdasarkan algoritma Tribase, dengan graph yang dibangun berdasarkan relasi retweet dimana retweet adalah suatu relasi interaksi antar user yang dikenal pada sosial media Twitter [6].

Dengan adanya penelitian sebelumnya, dapat dikatakan pada suatu bahasan topik dapat menghasilkan banyak data. Data yang dihasilkan tersebut dapat dilihat kemiripannya dan dianalisis lebih lanjut.

III. METODOLOGI DAN DATASET

A. Dataset

Pada penelitian ini objek dari tahapan *scraping* data adalah cuitan dari Twitter. Data yang akan *scraping* adalah cuitan yang menyebut terkait kampanye protocol Kesehatan pada

masa pandemik di DKI Jakarta. *Package* rtweet yang tersedia dalam R Studio digunakan untuk memudahkan pengambilan cuitan dari twitter yang biasanya membutuhkan Twitter API untuk mengaksesnya [7]. Pada scraping ini ada beberapa kata kunci yang menjadi sumber dari dataset, yaitu "protocol Kesehatan", "pakai masker", "cuci tangan", dan "jaga jarak". Dalam *scraping* setiap cuitan dihasilkan 90 atribut pada tiap *record* cuitan. Data yang dihasilkan dari *scraping* sebesar 15000 *record* dari seluruh kata kunci yang dicari. Dataset yang didapatkan ini hanya terdiri dari kumpulan *tweet* dengan kata kunci terkait selama seminggu kebelakang, dikarenakan akun Twitter yang dipakai memiliki batasan.

B. Preprocessing

Preprocessing data dilakukan dengan mengharapkan dapat menghindari data-data yang kotor dari dataset yang telah ditentukan. Dalam studi kasus ini beberapa langkah-langkah untuk melakukan preprocessing terhadap tweet suatu akun dilakukan seperti berikut:

1. Case folding. Merupakan tahap perubahan suatu huruf dari huruf kapital menjadi huruf kecil.
2. Penghapusan website, email dan simbol.
3. Penghapusan stopword. Stopword merupakan kata umum yang sering muncul dan dianggap tidak memiliki makna beberapa contoh stopword Bahasa Indonesia adalah "yang", "di", dan "ke".
4. Tokenisasi kalimat. Tokenisasi adalah proses memecah dokumen menjadi kumpulan kata. Tokenization dapat dilakukan dengan menghilangkan tanda baca dan memisahkannya per spasi.

C. Pengecekan Kemiripan *Tweet*

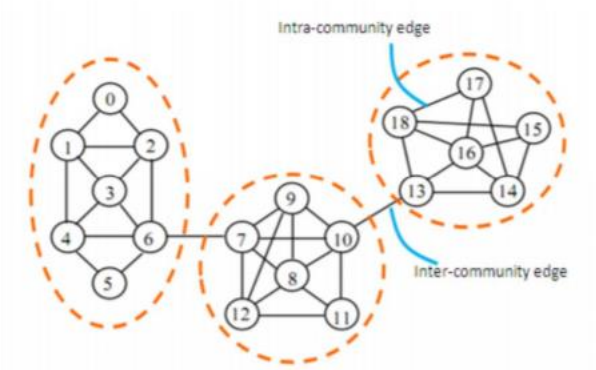
Metode TF-IDF Cosine Similarity bisa digunakan untuk menganalisa kesamaan atau kemiripan suatu dokumen teks dengan dokumen lainnya. Hal ini bisa digunakan untuk membandingkan suatu karya tulis, apakah plagiat atau bukan. Dan seberapa persen kemiripannya dengan karya tulis yang lain. TF-IDF digunakan untuk menghitung bobot suatu kata di dalam dokumen, sedangkan Cosine Similarity digunakan untuk pengecekan kemiripan dokumen tersebut. Rumus Cosine Similarity adalah sebagai berikut:

$$Cos\ \alpha = \frac{\mathbf{A} \bullet \mathbf{B}}{|\mathbf{A}||\mathbf{B}|} = \frac{\sum_{i=1}^{n} \mathbf{A_i} \times \mathbf{B_i}}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Dimana:
A = Vektor A, yang akan dibandingkan kemiripannya
B = Vektor B, yang akan dibandingkan kemiripannya
A • B = dot product antara vektor A dan vektor B
|A| = panjang vektor A
|B| = panjang vektor B
|A||B| = cross product antara |A| dan |B|

D. Deteksi Komunitas

Deteksi komunitas dalam graf bertujuan untuk menemukan suatu komunitas berdasarkan struktur jaringan. Beberapa node yang mirip dalam suatu graf akan membentuk suatu komunitas di dalam jaringan. Edge yang menghubungkan node dalam komunitas adalah Intra-community edge, sedangkan edge yang menghubungkan node dalam komunitas yang berbeda disebut Inter-community edge [8]. Gambar 1 menunjukkan Intra-community dan Inter-community edge antara komunitas yang berbeda.



Gambar 1 Struktur komunitas di dalam graf, dengan menunjukkan intra-community

Algoritma Louvain adalah salah satu algoritma unsupervised learning yang terbagi menjadi 2 fase yaitu: pengoptimalan modularitas dan agregasi komunitas. Setelah langkah pertama selesai, langkah kedua akan dieksekusi[10]. Keduanya akan terus berjalan hingga mencapai modularitas maksimal dan tidak ada perubahan di dalam graf.

IV. PERCOBAAN DAN HASIL

A. *Preprocessing*

Sebelum masuk kedalam pengecekan kemiripan tweet dilakukan tahap preprocessing dengan harapan mendapat informasi yang diperlukan dan membuang informasi yang tidak penting di dalam suatu tweet.

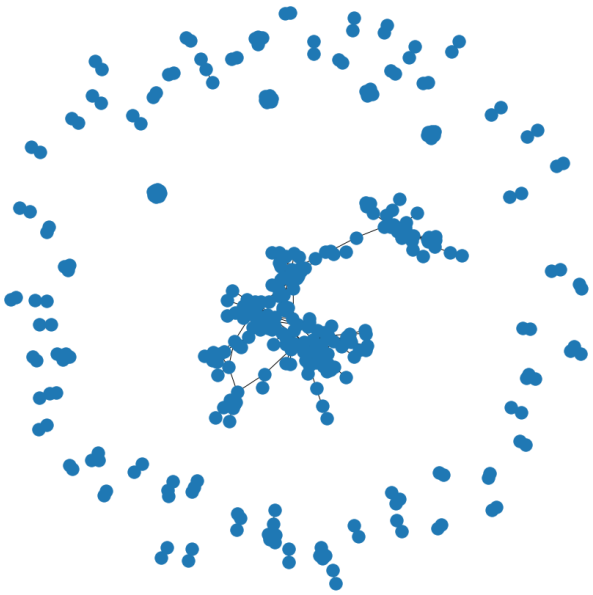
Sebelum Preprocessing	Setelah Preprocessing
'Kalau transit lama dalam perjalanan, ini hal harus diperhatikan:\r\n- ganti masker\r\n- rajin cuci tangan, cuci muka sekalian pakai sabun\r\n- kalau bisa mandi dan ganti baju, lebih baik lagi\r\n- makan dan minum vitamin agar imun tidak menurun\r\n- banyak minum air putih\r\n#TransmateJourney https://t.co/aZN8lteTIn'	transit perjalanan diperhatikan ganti masker rajin cuci tangan cuci muka pakai sabun mandi ganti baju makan minum vitamin imun menurun minum air putih

Tabel 1 Hasil preprocessing data

B. Kemiripan Tweet

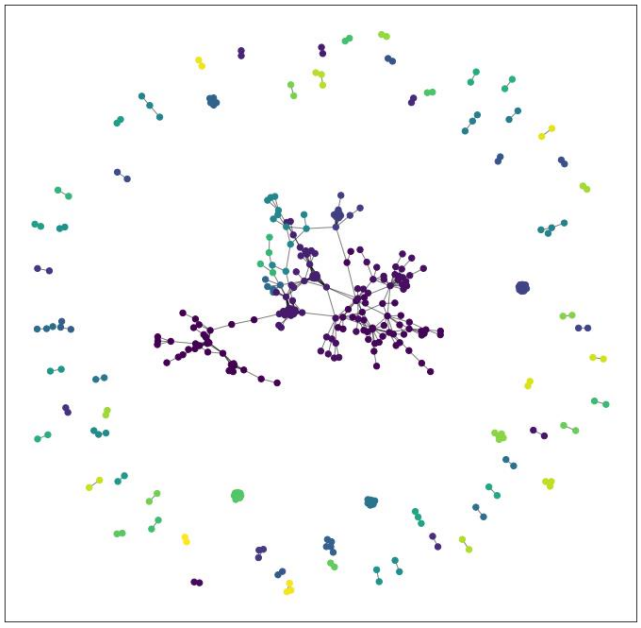
Setelah dilakukan *preprocessing* dilakukan pengecekan kemiripan *tweet* dengan metode TD-IDF Cosine Similarity.

Nilai kemiripan yang diambil adalah 60% sehingga menghasilkan graf seperti dibawah ini.



Gambar 2 Hasil Pengecekan Kemiripan

Setelah mendapatkan graf kemiripan, dilakukan deteksi komunitas dengan menggunakan logaritma Louvain dan dihasilkan bentuk graf komunitas seperti dibawah ini.



Gambar 3 Graf dari Komunitas

Komunitas yang terbentuk menghasilkan kumpulan kata-kata yang mirip dan dapat ditampilkan dalam *wordcloud* seperti berikut ini.

Gambar 4 Wordcloud dari Komunitas

V. KESIMPULAN DAN SARAN

Dari hasil dan analisis penelitian diatas dapat disimpulkan sebagai berikut:

1. Deteksi komunitas pada jejaring sosial media Twitter dapat dilakukan dengan algoritma Louvain.
2. Komunitas yang dihasilkan dapat memperlihatkan isu yang sedang dibicarakan di Twitter terkait protokol kesehatan.
3. Kebanyakan komunitas membahas tentang cuci tangan.

Pada penelitian selanjutnya, peneliti dapat mencoba menggunakan dataset dari media sosial yang berbeda seperti Facebook dan Instagram. Pada dataset yang sama penelitian dapat dilakukan lebih mendalam seperti medeteksi komunitas berdasarkan *sentiment* dari *tweet.*

VI. REFERENSI

[1] WHO, "Rolling updates on coronavirus disease (COVID-19)," WHO, 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen. [Accessed 20 September 2020].

[2] WHO, "Coronavirus," WHO, 2020. [Online]. Available:https://www.who.int/health-topics/coronavirus. [Accessed 20 September 2020].

[3] WHO, "Q&A on coronaviruses (COVID-19)," WHO, 17 April 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses. [Accessed 20 September 2020].

[4] K. K. R. Indonesia, "Keputusan Menteri Kesehatan Republik Indonesia Nomor HK.01.07/MENKES/413/2020 Tentang Pedoman Pencegahan dan Pengendalian Coronavirus Disease 2019 (COVID-19)," 2020.

[5] A. Irsyad and N. A. Rakhmawati, "Community detection in twitter based on tweets similarities in indonesian using cosine similarity and louvain

algorithms," Jurnal Ilmiah Teknologi Sistem Informasi (Scientific Journal of Information System Technology), vol. 1, pp. 22-32, 2020.

[6] Abdelsadek, Y., Chelghoum, K., Herrmann, F., Kacem, I. and Otjacques, B., “Community extraction and visualization in social networks applied to Twitter,” Information Sciences, vol. 424, pp.204-223, 2018.

[7] M. W. Kearney, "rtweet: Collecting and analyzing Twitter data," *Journal of Open Source Software*, vol. 42, no. 4, p. 1829, 2019.

[8] V. D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, "Fast unfolding of communities in large networks," J. Stat. Mech. (2008) P10008, p. 12, 2008.

[9] M. R. R. Gunaedi, I. Atastina and A. Herdiani, "Analisis dan Implementasi Algoritma Dynamicnet pada Deteksi Evolusi Komunitas di Media Sosial Twitter," in e-Proceeding of Engineering, 2018.