**Laporan Pengerjaan TK-2**

**Analitika Media Sosial**

**Kelompok E**

Arkha Sayoga Mayadi (1606890252/arkha.sayoga@ui.ac.id)

Muhammad Wildan Abdurrahman (1806191641/maulana.wildan@ui.ac.id)

Muhammad Jilham Luthfi (1806141340/muhammad.jilham@ui.ac.id)

Muhammad Ichsanul Amal (1906353454/muhammad.ichsanul91@ui.ac.id)

# **Daftar Isi**

[**Daftar Isi**](#_6cjqb2880vro) **2**

[**Tugas Kelompok 2**](#_p5bs6i1kc4kb) **3**

[**Data**](#_i9eagp8yrgi) **3**

[**Metodologi**](#_3l210zma9ae4) **6**

[Melengkapi dan Membersihkan Data](#_57yuq2e2ueif) 6

[Ekstraksi Fitur](#_rzty355ec64q) 6

[Pembelajaran Mesin](#_xluc29d1rdc8) 6

[***Research Question***](#_wvsz6y6aj24j) ***6***

[Nomor 1](#_6cw68hbpa0ou) 6

[Nomor 2](#_kb0iz9pkto4o) 8

[Nomor 3](#_yt9i5qzsvu7) 14

[Bonus Nomor 1](#_si0j6kb9qilh) 19

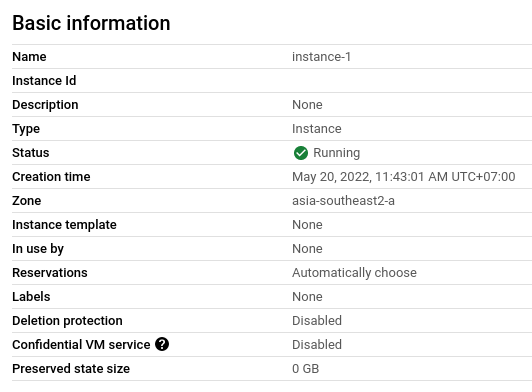
[Bonus Nomor 2](#_4zz1vq1py2t4) 22

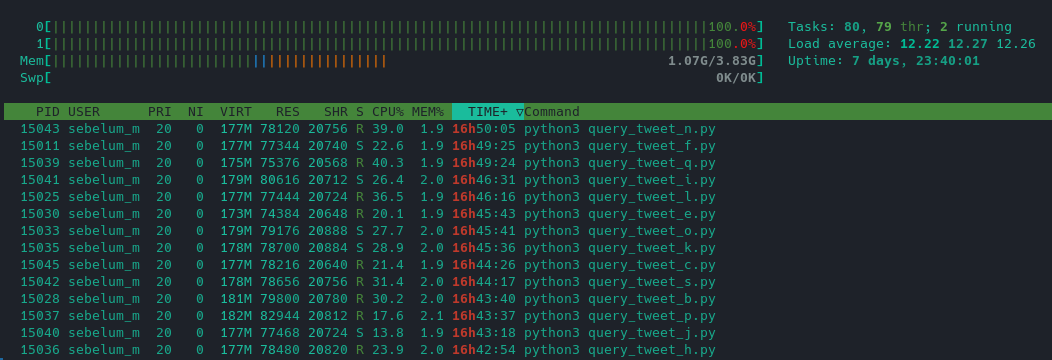
# Tugas Kelompok 2

Proyek ini merupakan mengerjakan proyek analitika media sosial *end-to-end* yang mencakup seluruh materi yang dipelajari selama perkuliahan satu semester, yakni *social media data collection*, *text mining*, *demography analysis*, dan *social network analytics*. Pada proyek ini, kami memilih *research question* 1a yang berkaitan dengan vaksinasi Covid-19. Berikut adalah pembahasan terkait yang telah kami lakukan untuk menjawab *research question* 1a.

# Data

Pengambilan data dilakukan dengan menggunakan tweet\_id dari [github.com/douglasraevan/covid19-indonesian-tweet](https://github.com/douglasraevan/covid19-indonesian-tweet). Filter dilakukan dengan tweepy untuk mendapatkan text tweet kemudian melakukan *query* ‘vaksin’. Dalam proses pengolahan data menggunakan teknik pemrograman paralel untuk mempercepat proses perolehan data. Dengan menjalankan ke 20 proses yang dijalankan di atas Google Compute yang kami buat untuk mengerjakan tugas ini. Dengan menggunakan 2 vCPUs dan ram 4GB program ini berjalan kurang lebih selama 7 hari.

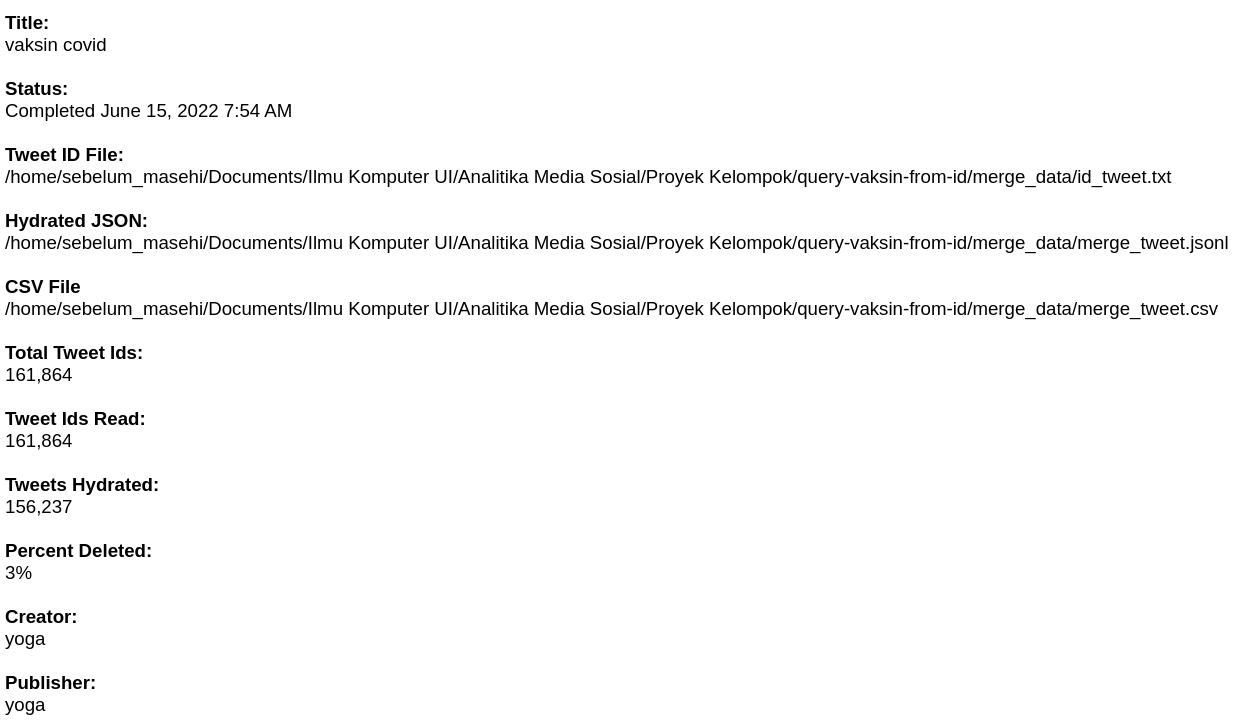
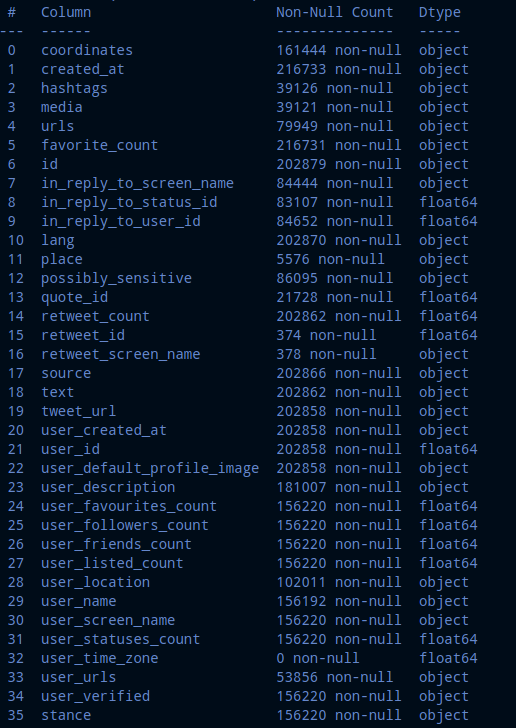
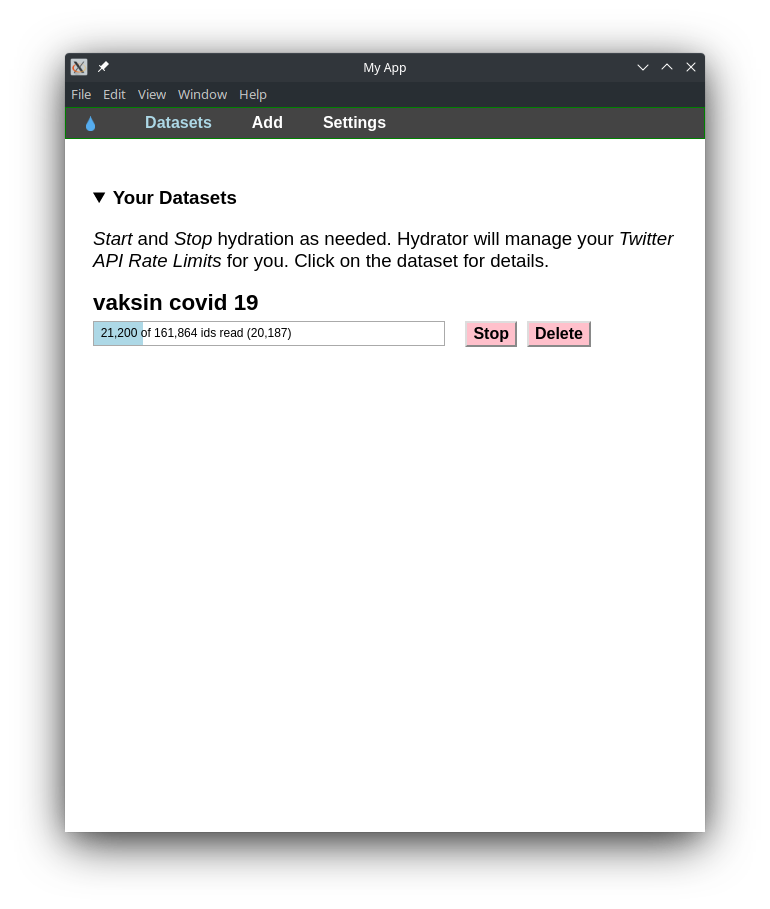




Pada proses pengumpulan data menggunakan *module* Twint, kami menentukan terlebih dahulu kata kunci pencarian yang akan digunakan. Kami menggunakan kata-kata, seperti vaksin, vaksinasi, dan beberapa produk vaksin yang beredar di Indonesia menurut laman resmi <https://covid19.go.id/tentang-vaksin-covid19>. Namun, kami tidak menggunakan kata janssen dan convidencia karena kami melihat tidak banyak menemukan *tweet* yang mengandung kata-kata tersebut. Hasil dari cara ini adalah kami mendapatkan 12.000 *tweet*. Kemudian, kami melihat beberapa sampel *tweet* dan menemukan ada *tweet* vaksin yang tidak berhubungan dengan covid, seperti hepatitis, cacar, kucing, kitten, dan anabul. Maka dari itu, kami melakukan filter kembali untuk menghilangkan *tweet* tersebut dan hasil akhir yang kami dapatkan adalah 9916 *tweet*.

Berdasarkan pertanyaan analitik yang dipilih, yakni RQ #1a, kami membuat anotasi data dengan label *stance* yang terdiri dari tiga label, yakni pro, kontra, dan abstain. Sebelum membuat label, kami berdiskusi singkat terkait kriteria dari ketiga label tersebut. Kami menetapkan bahwa *tweet* dengan label pro mengandung intensi berupa dukungan terhadap vaksin dan proses vaksinasi COVID-19. Sementara itu, *tweet* dengan label “KONTRA” mengadung intensi berupa penolakan terhadap vaksin dan proses vaksinasi COVID-19. Terakhir, *tweet* dengan label “ABSTAIN” menandakan tidak ada kecenderungan terkait vaksin, seperti pertanyaan terkait vaksin atau cerita pengalaman setelah vaksinasi.

Setelah mendapatkan semua id dan anotasi data, tahap selanjutnya kami melakukan teknik hidryrator dari [github.com/DocNow/hydrator](https://github.com/DocNow/hydrator). Dengan menggunakan hydrator ini akan mengembalikan semua data yang berkaitan dengan tweet hingga data tentang usernya.



Setelah mendapatkan seluruh data dari *hydrator* ini, kemudian melakukan penggabungan data antara hasil *hydrator* dengan anotasi.

# Metodologi

## Melengkapi dan Membersihkan Data

Selanjutnya, kami melakukan proses *preprocessing* pada data. Kolom yang akan kami gunakan adalah data *tweet*. Pada tahap *preprocessing*, kami menerapkan pembersihan data, seperti *lowercasing*, menghapus emoji, tanda baca, angka, dan proses *trimming*. Selain itu, kami mencoba untuk menghilangkan *stopwords* dan melakukan normalisasi dengan memperbaiki kata *slang* atau kata singkatan.

## Ekstraksi Fitur

Kami mentransformasikan data *tweet* menjadi bentuk numerik dengan vektor TF-IDF dengan ukuran 25.000 fitur berupa unigram hingga 4-gram.

## Pembelajaran Mesin

Selanjutnya, kami memilih model *multinomial naive bayes* karena model tersebut memberikan hasil terbaik pada data *tweet* berdasarkan proyek tengah semester. Kami juga melihat bahwa *task* terkait mendapatkan label dari sejumlah kecil data bisa menggunakan pendekatan *semi-supervised learning*. Teknik yang kami gunakan adalah *self-training*, yakni membangun model dari data berlabel untuk memprediksi data tidak berlabel. Kemudian, data dengan prediksi label yang tinggi akan dimasukkan ke dalam *training data* untuk pembangunan model berikutnya.

# *Research Question*

## Nomor 1

Verifikasi kebenaran pernyataan berikut dengan analisis data media sosial “Mayoritas orang yang bekerja di ranah kesehatan mendukung vaksinasi Covid-19”.

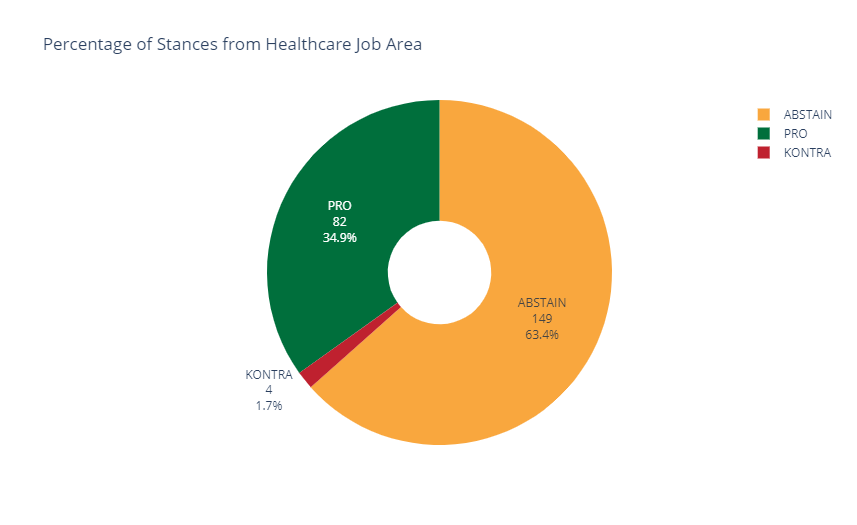
Untuk melakukan verifikasi kebenaran pernyataan di atas, kami perlu melakukan *demography analysis* terhadap *user* Twitter dengan memprediksi ranah pekerjaan mereka. Data sebelumnya yang dimiliki berupa data *tweet*. Kami mengambil sampel dengan teknik *stratified sampling* dari total 161.864 *tweet* yang berhasil dikumpulkan pada *checkpoint* sebelumnya, yakni 10% dari tiap kelas *stance*. Dari sampel tersebut, kami melakukan *random sampling* dengan mengambil sebanyak 2.500 pengguna karena keterbatasan komputasi.

Setelah mendapatkan data *username* pengguna, kami mengambil 20 *tweet* untuk masing-masing pengguna. Pada proses ini, kami menggunakan *module* Twint. Sayangnya, ada beberapa *tweet* yang tidak berhasil diambil sehingga kami membuang data *username* pengguna. Total ada 1.875 pengguna yang berhasil dikumpulkan *tweet*-nya. Dengan begitu, data tersebut bisa masuk ke proses berikutnya, yakni prediksi ranah pekerjaan.

| user\_tweet\_data = []  for username in tqdm(list\_username):  try:  c = twint.Config()  c.Pandas = True  c.Username = username  c.Limit = 15  c.Pandas = True  c.Hide\_output = True  twint.run.Search(c)  user\_tweet\_data.append(twint.storage.panda.Tweets\_df)    except ValueError:  user\_tweet\_data.append(pd.DataFrame()) |
| --- |

Sebelum memasuki model, kami melakukan pembersihan data *tweet* terlebih dahulu. Proses pembersihan melibatkan *lowercasing*, penghapusan kata “'b”, “RT”, emoji, kata berulang, dan spasi berlebih. Selain itu, kami juga menghapus *stopword* dan mengganti kata singkatan serta kata “*slang*”. Setelah *tweet* sudah dibersihkan, kami melakukan vektorisasi *tweet* menjadi representasi *tweet*. Model vektorisasi yang kami gunakan berasal dari Tugas Kelompok 1 milik Muhammad Jilham Luthfi karena menghasilkan model yang paling baik. Terakhir, kami melakukan prediksi dengan model dari Tugas Kelompok 1 milik Muhammad Jilham Luthfi juga. Akhirnya, setiap pengguna sudah memiliki atribut ranah pekerjaan.

Dalam menjawab pertanyaan nomor 1, kami menyaring kembali data dengan atribut “*job\_area*” bernilai “kesehatan” karena pada soal yang dibutuhkan adalah orang yang bekerja di ranah kesehatan. Kemudian, kami membuat visualisasi berupa *doughnut chart* seperti gambar di bawah ini. Terlihat bahwa **mayoritas pengguna pada ranah memiliki *stance* abstain** pada *tweet* mereka, yakni sebanyak 149 dari total 235 pengguna (63,4%). Sementara itu, orang yang pro terhadap vaksinasi Covid-19 sebanyak 82 pengguna (34,9%) dan orang yang kontra terhadap vaksinasi Covid-19 sebanyak 4 pengguna (1,7%). Dengan begitu, pernyataan bahwa “Mayoritas orang yang bekerja di ranah kesehatan mendukung vaksinasi Covid-19” **tidak benar**.



## Nomor 2

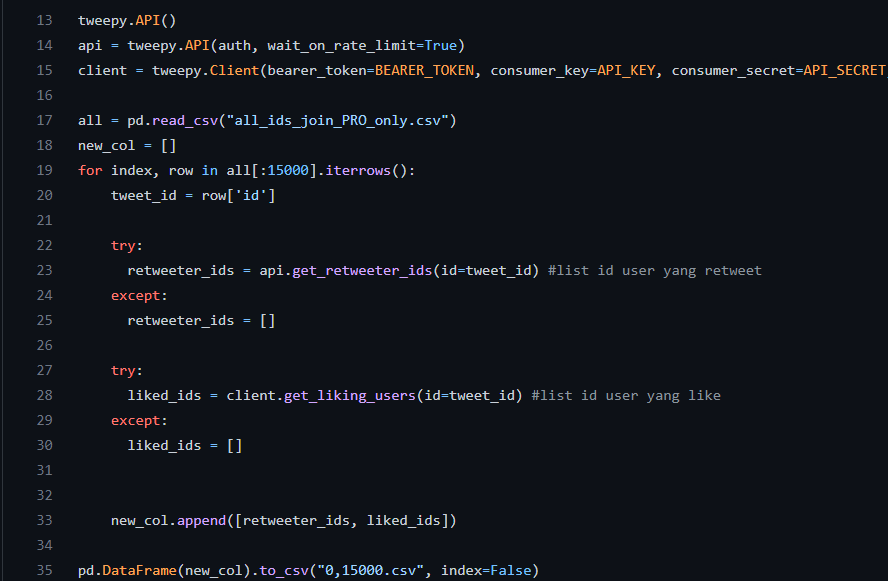
Analisis top 3 Twitter user yang memiliki pengaruh besar dalam mengkampanyekan vaksinasi Covid-19

Untuk dapat melakukan analisis top 3 Twitter *user* yang memiliki pengaruh besar dalam mengkampanyekan vaksinasi Covid-19, dilakukan pemilihan data terlebih dahulu. Data yang digunakan adalah data *tweet* yang memiliki *stance*, terdiri dari 161.864 baris. *Stance* dari *tweet* tersebut terdiri dari pro, kontra, dan netral, sehingga dilakukan pemilihan hanya *tweet* yang pro saja yang terdiri dari 42.863 baris. Kami asumsikan bahwa prediksi tersebut sudah benar berdasarkan tahapan sebelumnya. Selanjutnya, dilakukan *join* dengan data *tweet* berdasarkan id yang didapat dari Hydrator agar dataset *tweet* memiliki atribut yang lengkap.

Selanjutnya, kami mengeksplorasi definisi “memiliki pengaruh besar”. Dalam hal ini, “memiliki pengaruh besar” dapat dilihat dari jumlah *followers* dari pengguna yang mengunggah *tweet*, serta jumlah *retweet* dan jumlah *like* terhadap *tweet* yang diunggah pengguna. Selain itu, kami juga ingin melihat pengguna dengan *tweet* pro vaksinasi Covid-19 terbanyak. Berikut berbagai temuan kami:

| Pengguna dengan jumlah *followers* terbanyak |
| --- |
| Pengguna dengan jumlah *like* dan *retweet* terbanyak dari beberapa *tweet*-nya yang PRO vaksinasi Covid-19 |
| Pengguna dengan *tweet* PRO vaksinasi Covid-19 terbanyak |

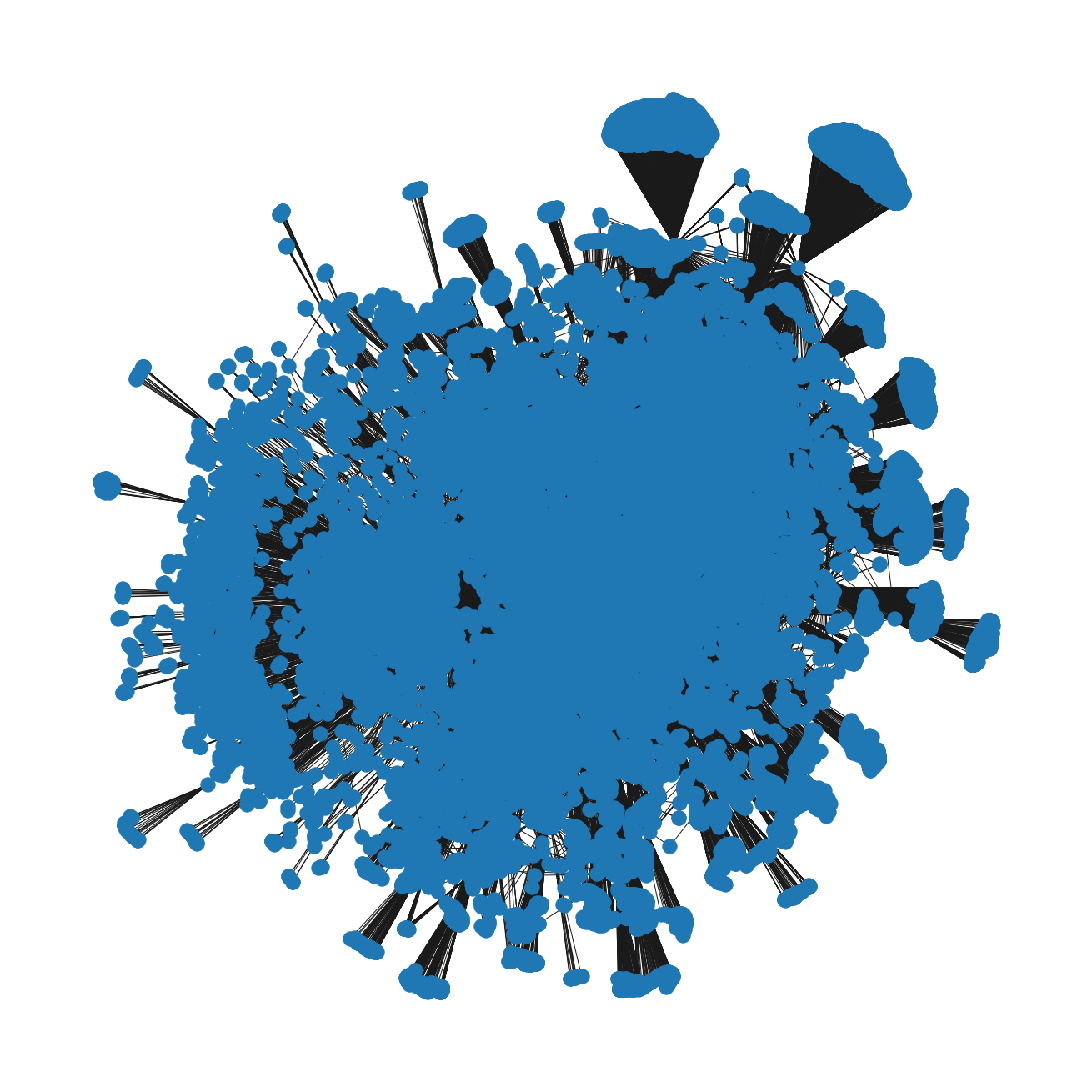
Setelah itu, dapat dilakukan juga *centrality analysis* berdasarkan jaringan antar pengguna dan aksinya yang melakukan *retweet* ataupun *like* terhadap *tweet* lainnya. Untuk itu, diperlukan data tambahan dengan fitur daftar id pengguna yang melakukan *retweet* terhadap *tweet*, dan juga id pengguna yang melakukan *like* terhadap *tweet*.



Dalam pengumpulan data tersebut, kami hanya mengambil 15 ribu data karena limitasi API dari Twitter yang dapat menghabiskan banyak waktu apabila ingin mengambil data lebih banyak lagi. Setelah mendapatkan data tersebut, dilakukan pembangunan *graph* dengan menggunakan *library* networkx dari Python.

*Graph* yang terbentuk adalah *multiplex graph* dimana *node* nya adalah id pengguna Twitter, dan *edge-*nya adalah suatu pengguna melakukan *retweet* atau *like* pada suatu *tweet* dari pengguna lainnya. Berikut merupakan rangkuman dan visualisasi dari *graph*.

* *Node*: User ID
* *Edge*:
  + User ID X melakukan retweet tweet dari user ID Y
  + User ID X melakukan like tweet dari user ID Y



Setelah *graph* berhasil dibuat, dijalankan algoritma *pagerank* untuk mengetahui pengguna mana yang paling sentral, dimana pengguna lainnya banyak melakukan *retweet* dan *like* yang menuju pada pengguna dari *tweet* tersebut. Kami menggunakan *pagerank centrality* untuk mengukur kepentingan *node* dari segi arus aliran **penyebaran informasi** (*information propagation*) berdasarkan *directed edge* pada graf.

Berikut merupakan pasangan *user* iddengan nilai *pagerank*-nya diurutkan dari yang terbesar:

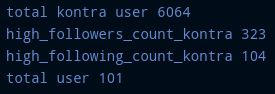
| ***User* ID** | ***Username*** | ***PageRank Centrality*** |
| --- | --- | --- |
| 69183155 | detikcom | 0.021899698828444693 |
| 17128975 | CNNIndonesia | 0.021518981880509505 |
| 71436318 | KompasTV | 0.009691609298089205 |
| 23343960 | kompascom | 0.009269979123187097 |
| 887743587579944960 | OposisiCerdas | 0.007717083211040062 |
| 161236190 | KemenkesRI | 0.0071091370838462856 |
| 1297362611026059269 | drpriono1 | 0.006418290620693366 |
| 759692754985242625 | kumparan | 0.006334692620596891 |
| 18129942 | tempodotco | 0.005749871988048712 |
| 41730943 | VIVAcoid | 0.0051038479392882564 |

Dapat dilihat bahwa 3 *user* yang memiliki pengaruh besar dalam mengkampanyekan vaksinasi Covid-19 berdasarkan algoritma *pagerank* adalah **detikcom**, **CNNIndonesia**, dan **KompasTV**. Interpretasi kami adalah ketiga akun tersebut merupakan akun media berita sehingga akun tersebut memiliki peran utama sebagai inisiator untuk menyebarkan berita tentang vaksinasi Covid-19 yang tentunya mendukung vaksinasi Covid-19. Selain itu, berdasarkan jumlah *like* dan *retweet* terbanyak, 3 *user* yang memiliki pengaruh besar dalam mengkampanyekan vaksinasi Covid-19 adalah **HausofHilton**, **nora\_mhhni**, dan **projekhitam**.

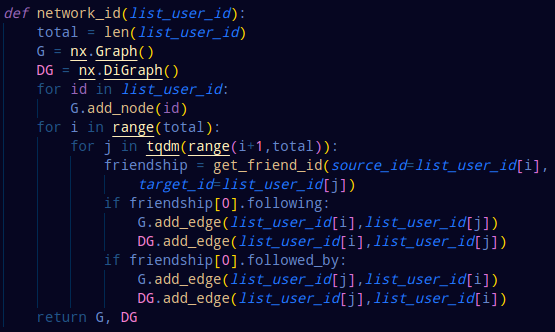
## Nomor 3

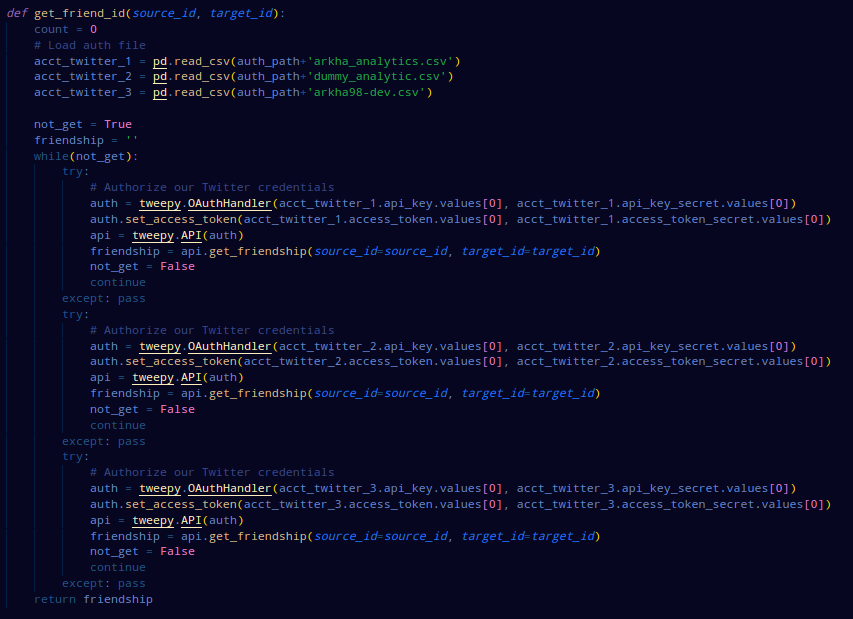
Analisis cluster pada social network user penentang vaksinasi Covid-19

Pengerjaan ini kami lakukan dengan melakukan pengambilan tweet yang kontra sesuai dengan proses anotasi yang telah dilakukan. Dari hasil pengambilan data dan anotasi yang dilakukan, ada 6356 tweet, dengan total user 6064. Oleh karena itu kami lakukan filter dari total 6064 user menjadi 101 user yang dipilih berdasarkan banyaknya total follower atau following terbanyak dalam hal ini diambil yang memiliki total following dan follower sebanyak 6000.

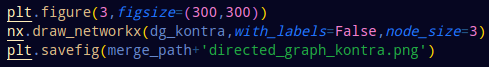


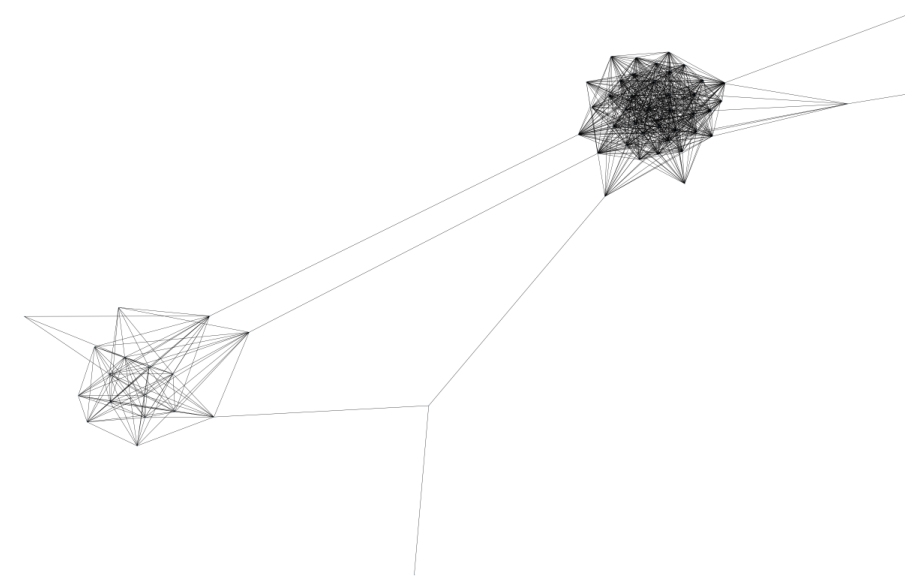
Setelah melakukan pengambilan user\_id kami melakukan pengambilan data get\_frienship dari user\_id menggunakan Tweepy. Karena ada 101 user maka total kemungkinan edge untuk undirected graph C(101,2) sedangkan untuk directed graph P(101,2).



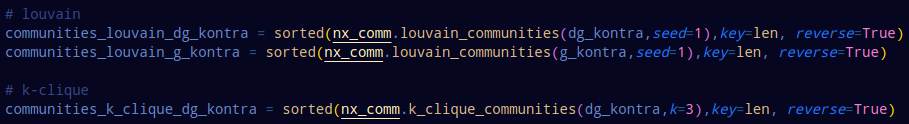


Setelah melakukan pengambilan get\_friendship dari semua kemungkinan edge maka menghasilkan bentuk graph.

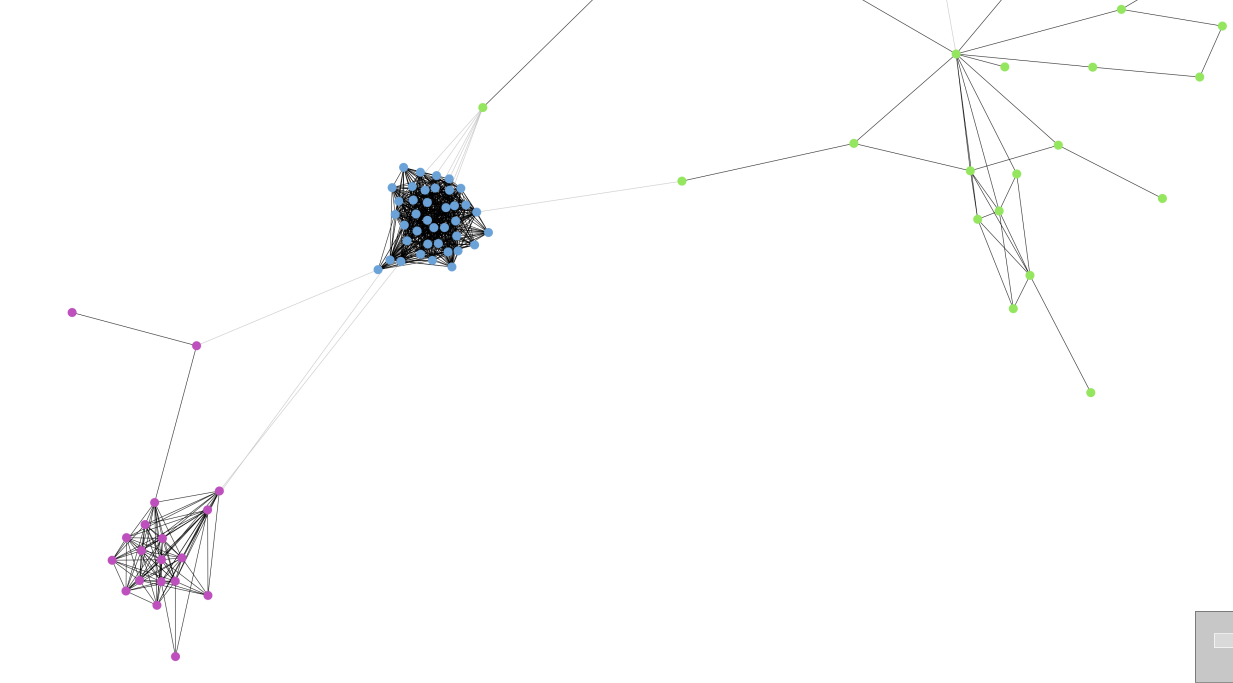




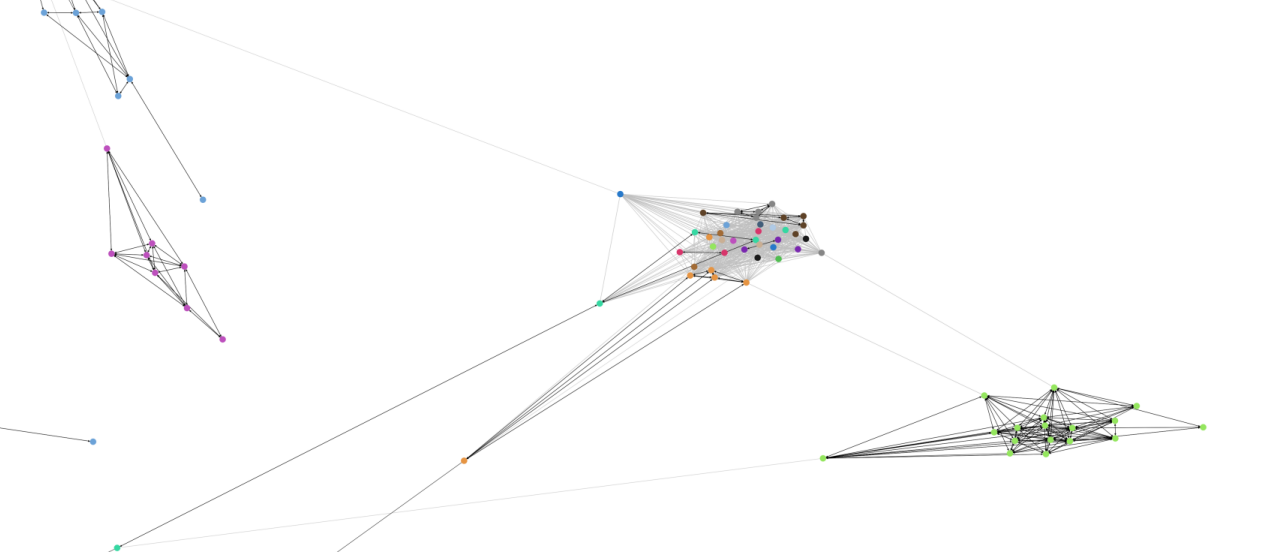
Untuk melakukan deteksi komunitas kami menggunakan 2 metode yaitu louvain dan k-clique dengan menggunakan dua jenis graph directed dan undirected graph.



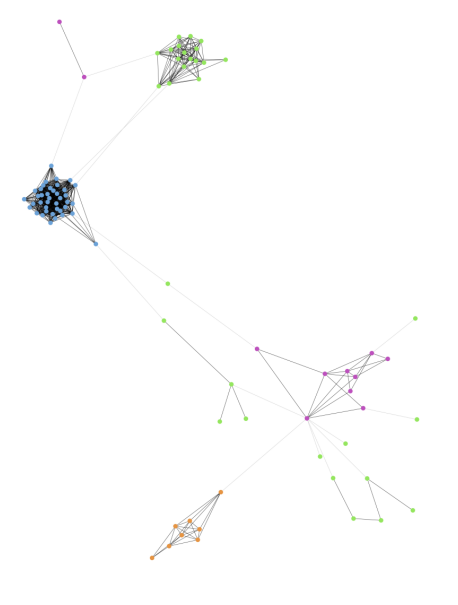
* Louvain
  + Directed Graph



* + Undirected Graph



* K-Clique
  + Directed Graph



Dari hasil deteksi komunitas ini didapatkan bahwa menggunakan directed graph memiliki deteksi komunitas yang lebih baik dibandingkan dengan undirected graph. Hal ini dikarenakan twitter lebih cocok dengan bentuk graph directed karena terdapat following (outgoing edge) dan followed\_by (incoming edge).

## Bonus Nomor 1

Vaksin Covid-19 apa saja yang mendapat sentimen positif (*stance* PRO) user masyarakat Indonesia?

Untuk menjawab pertanyaan ini, kami perlu mengetahui *tweet* mana saja yang termasuk sentimen positif. Sebelumnya kami telah memberikan label *stance* pada masing-masing tweet pada dataset dengan *stance* “PRO” untuk tweet yang mendukung vaksin atau memiliki sentimen positif. Oleh karena itu kami menggunakan dataset yang telah diberi label *stance* tersebut untuk menjawab pertanyaan ini. Kami menganalisis 10 jenis vaksin pada dataset ini sesuai dengan daftar jenis vaksin yang tertera pada website resmi pemerintah covid19.go.id (<https://covid19.go.id/tentang-vaksin-covid19>).

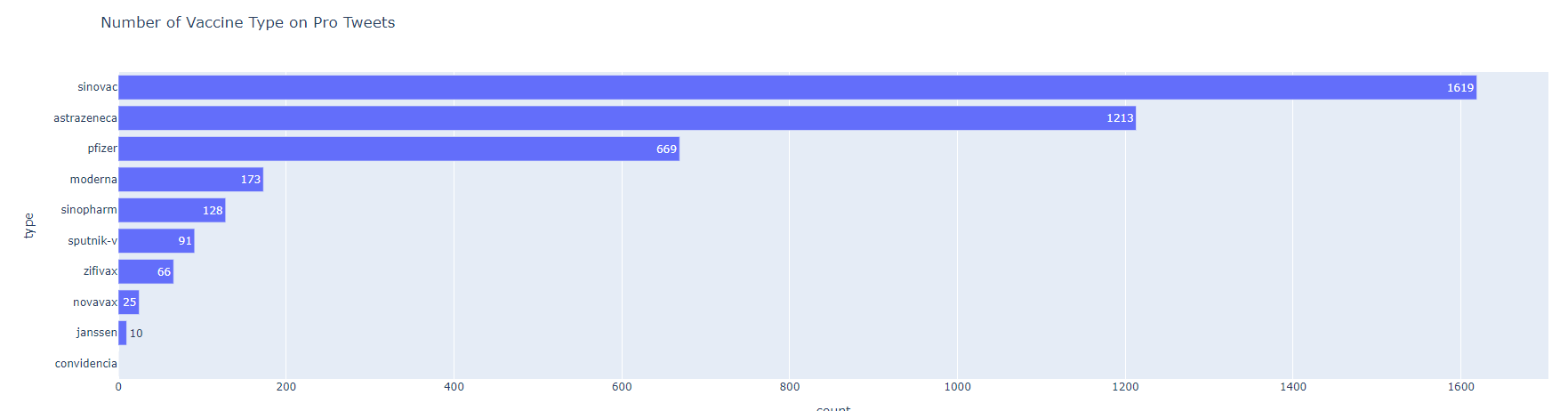
| # source: https://covid19.go.id/tentang-vaksin-covid19  vaccine\_type\_dict = {  'sinovac': ['sinovac'],  'astrazeneca': ['astrazeneca', 'astra', 'az'],  'sinopharm': ['sinopharm'],  'moderna': ['moderna'],  'pfizer': ['pfizer'],  'novavax': ['novavax'],  'sputnik-v': ['sputnik-v', 'sputnik'],  'janssen': ['janssen'],  'convidencia': ['convidencia'],  'zifivax': ['zifivax'],  } |
| --- |

| vaccine\_count\_dict = {key: 0 for key in vaccine\_type\_dict.keys()}  # for each tweet  for tweet in tqdm(clean\_tweet):  # check the occurence for each vaccine type by its common name in society  for vaccine, vaccine\_list in vaccine\_type\_dict.items():  # check for each common name  for vaccine\_type in vaccine\_list:  # count if exist and pass to another vaccine type  if vaccine\_type in tweet:  vaccine\_count\_dict[vaccine] += 1  break |
| --- |

Berdasarkan hasil dari 161.864 tweet yang diambil, **9 dari 10** jenis vaksin covid-19mendapatkan sentimen positif *user* (memiliki nilai *stance* “PRO”)*.* Berikut ini adalah jumlah tweet yang mengandung sentimen positif pada masing-masing jenis vaksin:

| vaccine\_count\_df = pd.DataFrame({  'type': list(vaccine\_count\_dict.keys()),  'count': list(vaccine\_count\_dict.values())  })  vaccine\_count\_df.sort\_values('count', inplace=True)  vaccine\_count\_df |
| --- |





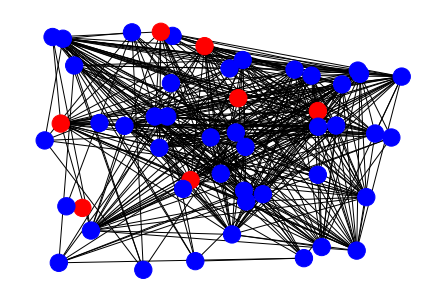
Dapat dilihat pada tabel dan grafik tersebut bahwa jenis vaksin yang paling banyak dibicarakan dan mendapat sentimen positif adalah **sinovac** dengan jumlah 1619 tweet diikuti dengan **astrazeneca** dengan 1213 tweet dan **pfizer** dengan 669 tweet. Sedangkan **convidencia** adalah jenis vaksin yang sama sekali tidak mendapat sentimen positif pada 161.864 sampel tweet yang kami analisis.

## Bonus Nomor 2

Apakah pertemanan (user di dunia maya) homogen berdasarkan sikap pro-kontra terhadap vaksin Covid-19 (misalnya apakah pro vaksin cenderung berteman dengan pro vaksin)?

Untuk menjawab pertanyaan bonus ini, kami memerlukan data pertemanan pengguna Twitter. Sebelumnya, kami telah mengumpulkan data tersebut untuk menjawab soal nomor 3. Data yang kami dapatkan sebelumnya berupa graf dengan *node* berupa pengguna dari semua kelas *stance*. Maka dari itu, kami menyaring kembali *node* yang hanya memiliki nilai *stance* “PRO” atau “KONTRA”. Kemudian, kami membangun kembali graf berdasarkan *node* yang sudah disaring dan *edge* sesuai graf sebelumnya. Informasi *stance* kami masukkan bersama dalam *node*. Hasil pembangunan graf bisa dilihat pada gambar di bawah ini. *Node* yang terbentuk berjumlah 50 buah dengan *node* berwarna biru menggambarkan pengguna yang pro terhadap vaksin Covid-19 dan *node* berwarna merah menggambarkan pengguna yang kontra terhadap vaksin Covid-19.

| # create empte graph  G = nx.Graph()  # insert node and its attribute (stance)  for \_, user in data\_user\_df.iterrows():  G.add\_node(user['user'], stance=user['stance'])  # insert edge (friendship)  for \_, friend in data\_friend\_df.iterrows():  G.add\_edge(friend['source'], friend['target']) |
| --- |



Berdasarkan graf di atas, kami menggunakan *homophily test* untuk menjawab pertanyaan bonus ini. Pertama-tama, kami menghitung jumlah *node* pengguna yang pro dan kontra terhadap vaksin Covid-19 sebagai nilai *p* (pro) dan *q* (kontra). Kemudian, kami menghitung jumlah *cross-edge* pada graf di atas. *Cross-edge* merupakan *edge* yang menghubungkan dua *node* dengan berbeda *stance*. Jumlah tersebut dibagi dengan jumlah total *edge* pada graf. Berdasarkan hasil perhitungan, ternyata nilai fraksi *cross-edge* (=**0.25**) melebihi nilai 2*pq* (≈**0.241**). Hal tersebut menandakan **belum ada cukup bukti** bahwa graf pertemanan pengguna di atas berdasarkan *stance* pro dan kontra terhadap vaksin Covid-19.

| # count pro user node and contra user node  node\_pro = 0  node\_kontra = 0  for node in list(G.nodes):  if G.nodes[node]['stance'] == 'PRO':  node\_pro += 1  elif G.nodes[node]['stance'] == 'KONTRA':  node\_kontra += 1  else:  continue |
| --- |
| # count cross edge from a graph  cross\_edge\_count = 0  for node, target in list(G.edges):  if G.nodes[node]['stance'] != G.nodes[target]['stance']:  cross\_edge\_count += 1 |
| # homophily test  threshold = 2 \* (node\_pro / len(G.nodes)) \* (node\_kontra / len(G.nodes))  cross\_edge\_prob = cross\_edge\_count / len(G.edges)  if cross\_edge\_prob < threshold:  print("Enough evidence to prove that there is a homophily in a graph")  else:  print("Not enough evidence") |