BAB III
PELAKSANAAN KERJA MAGANG

3.1 Kedudukan dan Koordinasi

Pada pelaksaan kerja magang di Blankspace, mahasiswa ditempatkan pada departemen *Tech & Product* sebagai *Data Engineer Intern*. Departemen tersebut melaksanakan beberapa pekerjaan seperti menerima permintaan *client* terkait pekerjaan di bidang *data*. Dalam mengerjakan tugas-tugas yang ada pada departemen tersebut, mahasiswa dibimbing oleh pembimbing lapangan yaitu Wandy Halim selaku *Data Scientist* Blankspace. Proyek penelitian dilaksanakan oleh 1 *Data Engineer Intern* dan 1 *Data Scientist*.

3.2 Tugas yang Dilakukan

Praktik kerja magang sebagai *Data Engineer Intern* di Blankspace terdapat beberapa tugas dan tanggung jawab yaitu mengumpulkan *data* yang akan digunakan oleh *Data Scientist* dalam melakukan pembuatan model. *Data Engineer* juga bertugas dalam *data cleaning* dan *preprocessing* agar *data* siap digunakan sebagai bahan baku pembuatan model *machine learning*. Selama 40 hari kerja, pekerjaan yang dilakukan adalah studi kasus analisis sentimen ujaran kebencian terhadap kelompok Ahmadiyah, Syiah, dan Cina di Indonesia. Topik tersebut dipilih secara langsung oleh pihak CSIS Indonesia dalam rangka adanya peningkatan ujaran kebencian yang berdampak pada tindakan / serangan fisik terhadap kelompok-kelompok minoritas. Seiring dengan peningkatan atau kemajuan teknologi terdapat perubahan bentuk ujaran kebencian yang sebelumnya secara *offline* / fisik menjadi *online* dengan menggunakan *platform* sosial media. Kelompok Cina, Ahmadiyah, dan Syiah dipilih berdasarkan *survey* CSIS, dimana ketiga kelompok tersebut merupakan kelompok minoritas yang mendapatkan paling banyak kekerasan fisik seperti pengusiran dan penyerangan yang diakibatkan oleh ujaran kebencian yang diarahkan kepada mereka. Tujuan dari *project* magang ini adalah untuk menyajikan *data* dan visualisasi mengenai 2 poin yaitu:

1) *Volume Hate Speech* terhadap kelompok Ahmadiyah, Syiah, dan Cina di Indonesia dengan rentang waktu hingga tahun 2024 dengan tujuan untuk mengetahui apakah ada peningkatan signifikan selama masa kampanye untuk menyambut pemilihan umum atau pemilu di masa yang akan datang.
2) Pemetaan aktor sosial yang masih berada dalam tahap lanjutan. Pemetaan aktor sosial ditujukan untuk mengetahui aktor sosial mana yang paling banyak berkontribusi dalam ujaran kebencian.

Dalam proyek magang ini menggunakan *data* yang diambil dari *Twitter* dalam bentuk *Tweet* dengan kata kunci 'Ahmadiyah', ' Syiah', dan 'Cina'. Uraian *timeline* kerja magang bisa dilihat pada tabel 3.1 yang menjelaskan tugas mahasiswa dalam melakukan kerja magang.

Tabel 3. 1 Tugas Kerja Magang

No.	Tugas yang Dilakukan	Mulai	Selesai
1.	**Riset teknik *scraping data* Twitter**		
1.a	Riset API Twitter	12/05/2021	13/05/2021
1.b	Riset *library* yang diperlukan dalam proses *scraping data*	13/05/2021	14/05/2021
2.	***Scraping Tweet* 'Syiah'**		
2.a	Melakukan pengambilan *data* atau *scraping* dengan kata kunci Syiah	17/05/2021	18/05/2021
2.b	Riset dan implementasi *regular expression* untuk *cleaning* pada hasil *scraping*	19/05/2021	20/05/2021
2.c	Riset *Natural Language Processing*	20/05/2021	21/05/2021
3.	***Scraping Tweet* 'Cina'**		

No.	Tugas yang Dilakukan	Mulai	Selesai
3.a	*Scraping data* Twitter dengan kata kunci Cina dan implementasi *regex* pada hasil *scraping*.	24/05/2021	25/05/2021
3.b	Menggabungkan semua *data* hasil *scraping* dan *labeling data* dengan 1 (mengandung ujaran kebencian) atau 0 (tidak mengandung ujaran kebencian).	26/05/2021	28/05/2021
4.	***Preprocessing data* 'Syiah' dan 'Cina'**		
4.a	*Filtering value* 1 dan 0.	31/05/2021	01/06/2021
4.b	*Merge* semua *data* dan menambahkan kolom baru terkait kata kunci (Syiah atau Cina).	02/06/2021	04/06/2021
5.	***Scraping Tweet* 'Cina' menggunakan *Geo location***		
5.a	Riset penggunaan *Geo location* dalam *scraping*	07/06/2021	08/06/2021
5.b	Melalukan *scraping* terkait *data Tweet* Syiah dengan kriteria *geo location* Jakarta, Indonesia	08/06/2021	09/06/2021
6.	***Push data* ke CSIS untuk di anotasi**		
6.a	*Prepare* semua *data* hasil *scraping*	09/06/2021	10/06/2021
6.b	*Split data* untuk diberikan kepada CSIS Indonesia agar dapat di anotasi	10/06/2021	11/06/2021
7.	***Clustering Tweet based on keyword***		
7.a	Riset mengenai model *machine learning*	14/06/2021	15/06/2021
7.b	Membuat *spreadsheet* terkait kata kunci *hate speech*	16/06/2021	18/06/2021

No.	Tugas yang Dilakukan	Mulai	Selesai
7.c	Mengkategorikan apakah *Tweet* tersebut mengandung ketiga kata kunci kolom "cindo" , "tkacina" , atau "*mainland*"	21/06/2021	22/06/2021
8.	***Scraping* dan *Preprocessing Tweet 'Ahmadiyah'***		
8.a	*Scraping* dan *Cleaning data* Twitter dengan kata kunci "Ahmadiyah" menggunakan *geo location* Jakarta, Indonesia	23/06/2021	25/06/2021
8.b	*Clustering Tweet* berdasarkan kata kunci "tka", "pki", "vaksin", "cindo"	28/06/2021	30/06/2021
9.	***Finalize Project***		
9.a	Membuat *code* baru yang lebih *simple* dan mudah dimengerti	01/07/2021	10/07/2021
9.b	*Finalize code* untuk digunakan oleh CSIS sebagai *reusable code*	11/07/2021	15/07/2021

3.3 Uraian Pelaksanaan Kerja Magang

Dalam pelaksanaan kerja magang di perusahaan Blankspace sebagai *Data Engineer Intern* dalam waktu 40 hari yang dimulai dari 12 Mei 2021 hingga 11 Agustus 2021 terdapat beberapa tahap yang dilakukan mahasiswa magang dalam berkontribusi pada proyek "Analisis Sentimen Ujaran Kebencian di Indonesia". Pada tahap pertama, mahasiswa akan melakukan proses *collecting data* atau *scraping data* melalui Twitter, selanjutnya tahap kedua melakukan *preprocessing data* pada hasil *output scraping*.

3.3.1 Collecting Data

Tugas yang dilakukan selama melakukan kerja magang adalah melakukan proses pengambilan *data* dan pembersihan *data* pada proyek penelitian ujaran kebencian di Indonesia. Dalam hal ini proyek tersebut

merupakan kerja sama antara kedua belah perusahaan yaitu blank-space.io dan CSIS Indonesia. Selama proyek tersebut, penulis menggunakan bahasa pemrograman python dan didukung oleh beberapa *library* seperti *Twint, numpy, pandas, datetime, nest_asyncio,* dan *library re.* Alasan digunakannya *library Twint* pada penelitian ini adalah dikarenakan *Twint* dapat mengekstrak *data* lebih dari 7 hari dengan *input* rentang waktu atau tanggal sesuai keinginan. *Data* yang diekstrak memiliki rentang waktu dari 1 Januari 2020 hingga 1 Juni 2021. Hasil dari *data* tersebut berupa csv dengan jumlah total 9000 *row* untuk topik Cina, 842 *row* untuk topik Syiah, dan 283 *row* untuk topik Ahmadiyah. Pada masing-masing hasil output berisi beberapa kolom yaitu *id, conversation_id, created_at, date, time, timezone, user_id, username, name, place, tweet, language, mentions, urls, photos, replies_count, retweets_count, likes_count, hashtags, cashtags, link, retweet, quote_url, video, thumbnail, near, geo, source, user_rt_id, user_rt, retweet_id, reply_to, retweet_date, translate, trans_src, trans_dest*. Semua kolom yang tertera pada csv adalah kolom yang dihasilkan berdasarkan *default scraping* menggunakan *library Twint*.

```
In [66]: pip install twint

         Requirement already satisfied: twint in /Users/gracevirnach/.local/lib/python3.8/site-packages (2.1.21)
         Requirement already satisfied: cchardet in /Users/gracevirnach/opt/anaconda3/lib/python3.8/site-packages (from twin
         t) (2.1.7)
         Requirement already satisfied: fake-useragent in /Users/gracevirnach/opt/anaconda3/lib/python3.8/site-packages (fro
         m twint) (0.1.11)
         Requirement already satisfied: pandas in /Users/gracevirnach/opt/anaconda3/lib/python3.8/site-packages (from twint)
         (1.1.3)
         Requirement already satisfied: aiohttp-socks in /Users/gracevirnach/opt/anaconda3/lib/python3.8/site-packages (from
         twint) (0.6.0)
         Requirement already satisfied: beautifulsoup4 in /Users/gracevirnach/opt/anaconda3/lib/python3.8/site-packages (fro
         m twint) (4.9.3)
         Requirement already satisfied: schedule in /Users/gracevirnach/opt/anaconda3/lib/python3.8/site-packages (from twin
         t) (1.1.0)
         Requirement already satisfied: geopy in /Users/gracevirnach/opt/anaconda3/lib/python3.8/site-packages (from twint)
         (2.1.0)
```

Gambar 3. 1 Install Library Twint

```
In [67]: pip install datetime

         Requirement already satisfied: datetime in /Users/gracevirnach/opt/anaconda3/lib/python3.8/site-packages (4.3)
         Requirement already satisfied: pytz in /Users/gracevirnach/opt/anaconda3/lib/python3.8/site-packages (from datetime
         ) (2020.1)
         Requirement already satisfied: zope.interface in /Users/gracevirnach/opt/anaconda3/lib/python3.8/site-packages (fro
         m datetime) (5.1.2)
         Requirement already satisfied: setuptools in /Users/gracevirnach/opt/anaconda3/lib/python3.8/site-packages (from zo
         pe.interface->datetime) (50.3.1.post20201107)
         Note: you may need to restart the kernel to use updated packages.

In [68]: pip install pandas

         Requirement already satisfied: pandas in /Users/gracevirnach/opt/anaconda3/lib/python3.8/site-packages (1.1.3)
         Requirement already satisfied: pytz>=2017.2 in /Users/gracevirnach/opt/anaconda3/lib/python3.8/site-packages (from
         pandas) (2020.1)
         Requirement already satisfied: python-dateutil>=2.7.3 in /Users/gracevirnach/opt/anaconda3/lib/python3.8/site-packa
         ges (from pandas) (2.8.1)
         Requirement already satisfied: numpy>=1.15.4 in /Users/gracevirnach/opt/anaconda3/lib/python3.8/site-packages (from
         pandas) (1.19.2)
         Requirement already satisfied: six>=1.5 in /Users/gracevirnach/opt/anaconda3/lib/python3.8/site-packages (from pyth
         on-dateutil>=2.7.3->pandas) (1.15.0)
         Note: you may need to restart the kernel to use updated packages.
```

Gambar 3. 2 Install Datetime dan Pandas

Pada gambar 3.1 dan gambar 3.2 menunjukkan beberapa *library* Python yang perlu diinstall dalam penelitian ini. Salah satunya adalah *library twint* yang digunakan sebagai Twitter *scraping tool* yang memungkinkan untuk *scraping Tweet* tanpa menggunakan API Twitter. Selain itu, terdapat beberapa *library* lain seperti *pandas, numpy, nest_asyncio, re,* dan *datetime*. Fungsi dari *library pandas* adalah untuk membantu dalam pembuatan struktur *data* serta analisis *data* yang dapat digunakan dalam pembuatan table dan mengubah dimensi *data*. Pada *library numpy* berguna dalam proses komputasi numerik pada bahasa pemrograman Python. *Library nest_asyncio* memiliki fungsi untuk membantu dalam proses *scraping* pada saat menggunakan *library twint*. Sedangkan library datetime membantu dalam menyediakan sejumlah fungsi yang berkaitan dengan waktu, tanggal, dan interval waktu.

Gambar 3. 3 Import Library

Gambar 3.3 menunjukkan beberapa *library* yang telah di *install* sebelumnya dan perlu di *import* dalam penelitian ini. *Library* tersebut diantaranya adalah *twint, datetime, pandas, numpy, nest_asyncio,* dan *re.*

```
In [6]: current_date = datetime.datetime(2020,1,1)
        current_end_date = current_date + datetime.timedelta(days=1)
        end_date = datetime.datetime(2021,6,1)
        while (current_date != end_date):
            c = twint.Config()
            c.Search ='cina'
            c.Limit = 1000
            c.Lang = 'in'
            c.Geo ='-6.185543, 106.844082,500km'
            c.Store_csv = True
            c.Output = 'OutputCina.csv'
            c.Since = current_date.strftime("%Y-%m-%d")
            c.Until = current_end_date.strftime("%Y-%m-%d")
            twint.run.Search(c)
            current_date = current_end_date
            current_end_date += datetime.timedelta(days=1)
        print(current_date != end_date)
```

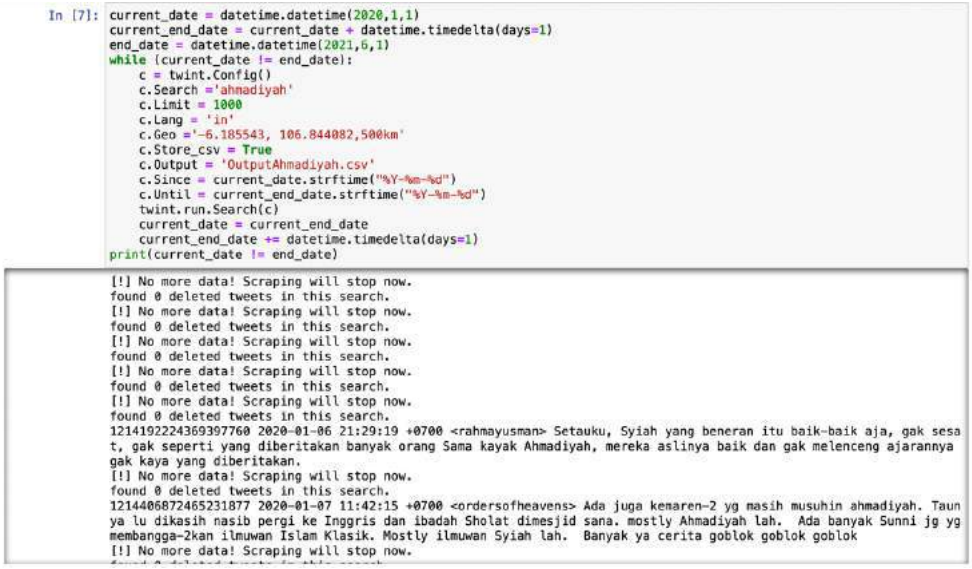
```
1212475813926322176 2020-01-02 03:48:55 +0700 <lidbahaweres> @CommuterLine  pagi ini dari Pondok Cina ke Tanah Aban
g jam 5 pagi apakah sudah normal? Tq
1212463631025111040 2020-01-02 03:00:30 +0700 <Iiefachri> @liem_id Jd hongkong bkn cina ya.. br tau
1212463131328299008 2020-01-02 02:58:31 +0700 <dhikamardiyya> India, cina, medan, manado, ntt, depok, pwk, bogor, d
an jkt kumpul dsini.  https://t.co/Whn3CxpNwY
1212451736641605633 2020-01-02 02:13:14 +0700 <boypramudyafana> @ACTforHumanity @Ajeng__Cute16 Kelurahan bidara cin
a kec. Jati negara jaktim. Saat ini masyarakat yg terdampak banjir mengungsi di Gor Yusenter otista raya.  https://
t.co/FiS2daDk2w
1212421432732340225 2020-01-02 00:12:49 +0700 <syahriells> @yarraPD Ya allah, pondasinya ga kokoh, soalnya bukan or
anh cina
1212421373219373056 2020-01-02 00:12:35 +0700 <ghatoot> @joesvult @fiqiehaqie Matane nan ra kui? Mesti bidara cina
i? 😅
1212415677270618112 2020-01-01 23:49:57 +0700 <amalmanda_> Rupa rupi jalmi mah. Saur barudak mah cina natepan hajat
, atanapi istikhoroh.  Udh ditinggal bacain yasin biar arwahnya tenang Iyah gak @ErvinaRizqi1
1212414648722452483 2020-01-01 23:45:52 +0700 <LawuSusanto> @GesuriID @evndari Dukung dan usir kapal2 cina yg masuk
wilayah NKRI.
1212410069435813888 2020-01-01 23:27:40 +0700 <toekangkuliner> Bidara cina air sampai meluap
1212398224100614145 2020-01-01 22:40:36 +0700 <shiningcrush_> Ya allah alhamdulillah dah mulai banyak rute cina dan
beberapa ada yg layover sana nih tq perusahaanku makin mendekatkanku pada idol
```

Gambar 3. 4 Scraping Data Twitter Keyword ‘Cina’

```
In [2]: current_date = datetime.datetime(2020,1,1)
        current_end_date = current_date + datetime.timedelta(days=1)
        end_date = datetime.datetime(2021,6,1)
        while (current_date != end_date):
            c = twint.Config()
            c.Search ='syiah'
            c.Limit = 1000
            c.Lang = 'in'
            c.Geo ='-6.185543, 106.844082,500km'
            c.Store_csv = True
            c.Output = 'OutputSyiah.csv'
            c.Since = current_date.strftime("%Y-%m-%d")
            c.Until = current_end_date.strftime("%Y-%m-%d")
            twint.run.Search(c)
            current_date = current_end_date
            current_end_date += datetime.timedelta(days=1)
        print(current_date != end_date)
```

```
1212448121965297664 2020-01-02 01:58:53 +0700 <fajarnugros> @patriaguides @rohmatsyahru @arieparikesit Di rombongan
ada org Syiah juga, nah dia pengennya lepas juga utk ziarah sendiri, nah saya jadi kepo kan? Apa yg diziarahin? Den
ger cerita sejarah dll. Kejayaan Ottoman dll. Makanya saya pengen ke Turki.
1212248153862225920 2020-01-01 12:44:17 +0700 <alrasyidstore> @dyla_se @SKlaMud @yusuf_dumdum Jebule si syiah @yusu
f_dumdum kakeane tenan.
[!] No more data! Scraping will stop now.
found 0 deleted tweets in this search.
1212735003982827521 2020-01-02 20:58:51 +0700 <Burhanudin1994> @Restichayah @HusinShihab Yg dipikirin si syiah @Hus
inShihab cmn selangkangan makannya gk bs membadakan wilayah dki dan banten
1212549845094789120 2020-01-02 08:43:05 +0700 <AlGhuyub> @dedeindry_0304 Yang jelas dia pro zionis israel bantai ra
kyat palestine, pro cina komunis dzalim ke rakyat Uyghur dan pro Syiah
[!] No more data! Scraping will stop now.
found 0 deleted tweets in this search.
1213246294912729089 2020-01-04 06:50:32 +0700 <EPras92> @mcflurrymdd @RizPrima Lebelisasi Syiah sesat, bukan Islam,
teroris dll itu juga tdk lepas dari kepentingan AS dkk.
1213125002360344576 2020-01-03 22:48:33 +0700 <bayuanggara08> @akhaidhir Udah China koalisi sama Rusia udah deh kom
unis kumpul, tambah Iran Syiah ga disukai banget sama Sunni. Ya udah mumpung lawan jadi satu.  Pada akhirnya Romawi
(AS Israel dkk) &amp;amp; Muslim yg bersatu menang. Tapi keduanya bertempur lagi pake panah dan pasukan kuda dipimpin M
ahdi.
```

Gambar 3. 5 Scraping Data Twitter Keyword ‘Syiah’



Gambar 3. 6 Scraping Data Twitter Keyword 'Ahmadiyah'

Gambar 3.4 menunjukkan adanya proses *scraping* pada *data Tweet* Cina dan *scraping data* pada *Tweet* Syiah terdapat pada Gambar 3.5 serta *scraping data Tweet* Ahmadiyah terlihat pada Gambar 3.6. *Scraping* dilakukan dengan menggunakan parameter waktu dan kata kunci sesuai topik seperti Cina, Ahmadiyah, atau Syiah. Pengambilan *data* dilakukan pada saat *code scraping* di *running* sehingga *data* yang didapat tidak *real-time*. Parameter *date* dalam *code* Gambar 3.4 merepresentasikan tanggal awal hingga tanggal akhir dari pengambilan *data*. Jika diperlukan *data* tambahan atau *data* baru dengan rentang waktu yang berbeda, maka dapat mengganti parameter *date* sesuai kebutuhan. *Data* baru yang telah di *scraping* dapat digabungkan dengan *data* sebelumnya menggunakan fungsi *concat* pada bahasa pemrograman python. Selain itu, *Geo location* pada proses *scraping* berguna untuk mengambil *data* yang hanya berwilayah di Indonesia agar *data Tweet* yang diambil *valid* untuk merepresentasikan *Tweet Hate Speech* di Indonesia.

```
In [92]: df = pd.read_csv(r"OutputCina.csv")
         df.sample(20)
```

Out[92]:

	id	conversation_id	created_at	date	time	timezone	user_id	username	name	place	...	
6443	1313077080045744128	1313077080045744128	2020-10-05 18:22:26 WIB	2020-10-05	18:22:26	700	1005842250	hellennecs	(｡ ˘ д)ﾉ STOP!	NaN	...	-6.1855
8205	1359840727878885377	1359840727878885377	2021-02-11 19:24:28 WIB	2021-02-11	19:24:28	700	108591210	yoshu_sudarso	Yoshu Sudarso	NaN	...	-6.1855
5903	1300068420625395718	1300067568376397825	2020-08-30 20:50:40 WIB	2020-08-30	20:50:40	700	132435339	arddhe	a	NaN	...	-6.1855
9217	1394954604207644673	1394954604207644673	2021-05-19 16:54:29 WIB	2021-05-19	16:54:29	700	180067095	ericksadewa01	Erick Sadewa	NaN	...	-6.1855
5912	1299927086837018625	1299920001286037506	2020-08-30 11:29:03 WIB	2020-08-30	11:29:03	700	1034345893837688832	prasety72595950	Prasetyo	NaN	...	-6.1855
7321	1335093506663096321	1335054366714171394	2020-12-05 12:27:51 WIB	2020-12-05	12:27:51	700	1206215455657955328	aditwongs	aaaaadit	NaN	...	-6.1855
7964	1351829563039379457	1351829563039379457	2021-01-20 16:50:58 WIB	2021-01-20	16:50:58	700	1293114684657111041	dinanur64826836	Dina Nurdiana	NaN	...	-6.1855

Gambar 3. 7 Dataframe Output Topic Cina

```
In [3]: df = pd.read_csv(r"OutputSyiah.csv")
        df.sample(20)
```

Out[3]:

	id	conversation_id	created_at	date	time	timezone	user_id	username	name	place
596	1342280892228673536	1342272591319367680	2020-12-25 08:27:57 WIB	2020-12-25	08:27:57	700	747725422339698688	okikasepp	ki	NaN
594	1342301983621873664	1342129513577078786	2020-12-25 09:51:46 WIB	2020-12-25	09:51:46	700	136887216	fire_lotus	Catatan dari lereng Merbabu	NaN
637	1346219846330028033	1345744941859803136	2021-01-05 05:19:57 WIB	2021-01-05	05:19:57	700	963751550	kangsemproel	Hilik Ku Aink Lah#	NaN
300	1263408106857324544	1263250796692439040	2020-05-21 16:55:40 WIB	2020-05-21	16:55:40	700	3849043754	jackhan32768669	wowww	NaN
758	1392764371504308226	1392759114359468032	2021-05-13 15:51:17 WIB	2021-05-13	15:51:17	700	1070614305106649088	yanti_9294	Yanti_R	NaN
723	1379854394922045444	1379854394922045444	2021-04-08 00:51:38 WIB	2021-04-08	00:51:38	700	355196529	abdullah_assg	#FreePalestine	NaN
484	1308562892212461568	1308559469819899905	2020-09-23 07:24:40 WIB	2020-09-23	07:24:40	700	1253476171	safirasindarti	fira	NaN
230	1240159988477235200	1240159988477235200	2020-03-18 13:15:56 WIB	2020-03-18	13:15:56	700	146039313	abinyasalma	ابو سلمى الاندونيسي	{'type': 'Point', 'coordinates': [-6.34948189,...	
612	1343044614500929537	1343044614500929537	2020-12-27 11:02:43	2020-12-27	11:02:43	700	1115514215350054912	fireman3482	Fireman348	NaN

Gambar 3. 8 Dataframe Output Topic Syiah

```
In [13]: df = pd.read_csv(r"OutputAhmadiyah.csv")
         df
```

Out[13]:

	id	conversation_id	created_at	date	time	timezone	user_id	username	name	place	...
0	1214192224369397760	1214192224369397760	2020-01-06 21:29:19 WIB	2020-01-06	21:29:19	700	517170383	rahmayusman	Rahmah	NaN	...
1	1214406872485231877	1214405610562736128	2020-01-07 11:42:15 WIB	2020-01-07	11:42:15	700	185402933	ordersofheavens	Mvly▲bvdi	NaN	...
2	1218016588609269761	1218016588609269761	2020-01-17 10:45:59 WIB	2020-01-17	10:45:59	700	732263036	kata_dhoni	ali romdhoni	{'type': 'Point', 'coordinates': [-6.4890627, 106.7391402]}	...
3	1218755223713284096	1218755223713284096	2020-01-19 11:41:03 WIB	2020-01-19	11:41:03	700	732263036	kata_dhoni	ali romdhoni	{'type': 'Point', 'coordinates': [-6.90322436, 108.49546298]}	...
4	1219184791825731584	1219184791825731584	2020-01-20 16:08:00 WIB	2020-01-20	16:08:00	700	97331907	asmutaqi	AS MUTAQI #S4LAM	NaN	...
...	...	...	...	...	...	...	...	...	...	...	...
281	1393543416525459464	1393267290053898240	2021-05-15 19:26:55 WIB	2021-05-15	19:26:55	700	989881481848766465	mkaikabayoran	Majlis Khuddam-ul-Ahmadiyyah Kebayoran	NaN	...
282	1394265376498749441	1394265372094701569	2021-05-17 19:15:44 WIB	2021-05-17	19:15:44	700	304885479	azhrtfqrhmn	jaarm	NaN	...
283	1394884269911134211	1394884269911134211	2021-05-19 12:15:00 WIB	2021-05-19	12:15:00	700	2183263291	mypst_	Yusuf	NaN	...

Gambar 3. 9 Dataframe Output Topic Ahmadiyah

Gambar 3.7 menunjukkan proses pembuatan *dataframe* berdasarkan *output* hasil *scraping* pada topik Cina sebelumnya yang masih berbentuk csv. Gambar 3.8 dan Gambar 3.9 menunjukkan *dataframe* hasil *output* dari *scraping* sebelumnya dengan topik Syiah dan Ahmadiyah. Pada *dataframe* tersebut belum dilakukan proses *preprocessing* sehingga *data* yang terdapat dalam *dataframe* masih bersifat *data* mentah yang belum dilakukan perubahan atau pembersihan.

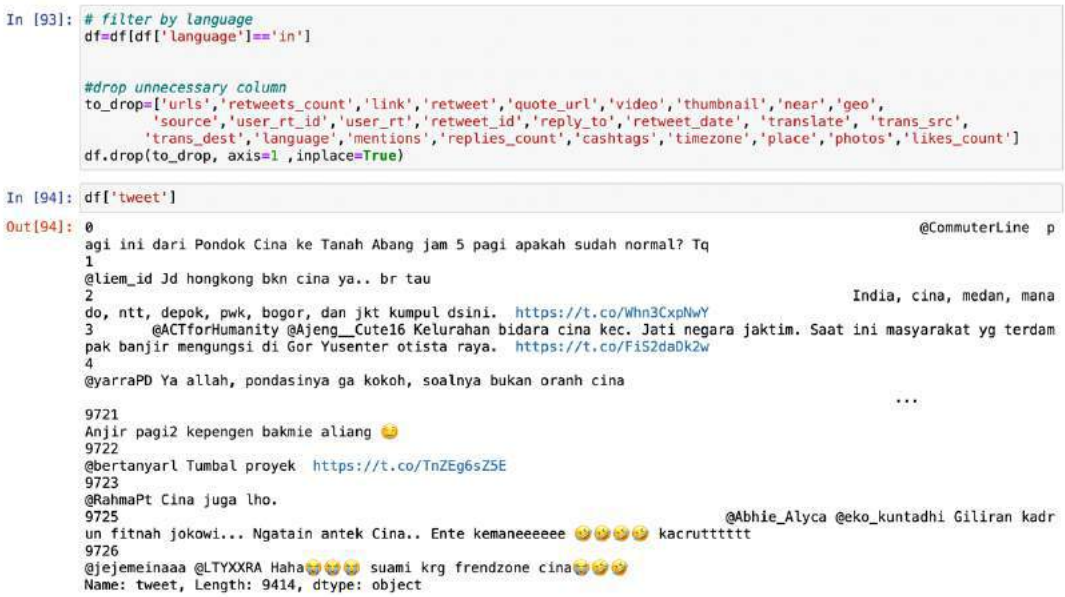
3.3.2 Data Preprocessing

Pada bagian pembersihan *data* menggunakan beberapa *library* lainnya untuk mengolah *data* dalam bentuk *dataframe*. *Data* yang diolah berupa hasil ekstrak *data* dari Twitter yang berbentuk csv dan ditransformasi ke dalam bentuk *dataframe* menggunakan bahasa pemrograman Python. Pada penelitian ini mengekstrak *data Tweet* yang berkaitan dengan 3 kelompok yang telah dipilih oleh CSIS yaitu kelompok Cina Indonesia, Syiah dan Ahmadiyah. Pembersihan *data* dilakukan dengan mengolah *data* menggunakan *regex* atau *regular expression* untuk membersihkan suatu kalimat pada *Tweet* seperti *mentions, number, hashtag, symbol, underscore, new line*, dan *retweet word*. Tabel 3.2 menunjukkan

adanya penambahan kolom pada *dataframe* yang bertujuan untuk mempermudah dalam proses anotasi sebelum dilakukan pembuatan model *machine learning*.

Tabel 3. 2 Add New Column to Dataframe

Nama	Keterangan
Uncleaned Tweet	*Tweet* yang belum dibersihkan sebagai cadangan jika diperlukan *Tweet* yang asli.
Topic	Terdiri dari China, Syiah, Ahmadiyah
Hate Speech	1 yang berarti mengandung ujaran kebencian, 0 yang berarti tidak mengandung ujaran kebencian.
Words Count	Jumlah kata dalam *Tweet*.
Contains link	Apakah *Tweet* tersebut mengandung link, 1 yang berarti iya, 0 yang berarti tidak
Category	Terdapat beberapa kategori yang ada yaitu TKA, *Mainland* China, Cina Indo, Vaksin, dan PKI.



Gambar 3. 10 Drop Column Cleaning Dataframe

Gambar 3.10 menunjukkan proses *cleaning dataframe* dengan menghapus beberapa kolom yaitu *urls, retweet_count, link, retweet, quote_url*, *video, thumbnail, near, geo, source, user_rt_id, user_rt, retweet_id, reply_to, retweet_date, translate, trans_src, trans_dest, language, mentions, replies_count, cashtahs, timezone, place, photos,* dan *likes_count*. Kolom yang dihapus merupakan kolom yang kurang memiliki korelasi dengan variabel utama yang akan diteliti yaitu ujaran kebencian melalui *Tweet*.

```
In [95]: pd.set_option('display.max_colwidth', None)
         df['topic']= 'Cina'
         df['hatespeech']= None
         df
```

Out[95]:

d	conversation_id	created_at	date	time	user_id	username	name	tweet	hashtags	topic	hatespeech
6	1212475813926322176	2020-01-02 03:48:55 WIB	2020-01-02	03:48:55	146038671	lidbaheweres	Rach Alida Bahaweres	@CommuterLine pagi ini dari Pondok Cina ke Tanah Abang jam 5 pagi apakah sudah normal? Tq	[]	Cina	None
0	1212178324602011648	2020-01-02 03:00:30 WIB	2020-01-02	03:00:30	1243362931	iiafachri	fachri52	@liem_id Jd hongkong bkn cina ya.. br tau	[]	Cina	None
8	1212463131328299008	2020-01-02 02:58:31 WIB	2020-01-02	02:58:31	210339049	dhikamardiyya	dhik	India, cina, medan, manado, ntt, depok, pwk, bogor, dan jkt kumpul dsini. https://t.co/Whn3CxpNwY	[]	Cina	None
3	1212185100038721537	2020-01-02 02:13:14 WIB	2020-01-02	02:13:14	960287510319448065	boypramudyafana	Sabriel Rayyan	@ACTforHumanity @Ajeng__Cute16 Kelurahan bidara cina kec. Jati negara jaktim. Saat ini masyarakat yg terdampak banjir mengungsi di Gor Yusenter otista raya. https://t.co/FI52daDk2w	[]	Cina	None
5	1212420404800708609	2020-01-02 00:12:49 WIB	2020-01-02	00:12:49	70074012	syahriells	Syahril Sidik	@yarraPD Ya allah, pondasinya ga kokoh, soalnya bukan oranh cina	[]	Cina	None

Gambar 3. 11 Add Column ‘Topic Cina’ and ‘Hate Speech’

```
In [6]: pd.set_option('display.max_colwidth', None)
        df['topic']= 'Syiah'
        df['hatespeech']= None
        df
```

Out[6]:

id	conversation_id	created_at	date	time	user_id	username	name	tweet	hashtags	topic	hatespeech
65297664	1212357552685514758	2020-01-02 01:58:53 WIB	2020-01-02	01:58:53	36321448	fajarnugros	Balada Si Roy - Segera!	@patriaguides @rohmatsyahru @ariieparikesit Di rombongan ada org Syiah juga, nah dia pengennya lepas juga utk ziarah sendiri, nah saya jadi kapo kan? Apa yg diziarahin? Denger cerita sejarah dll. Kejayaan Ottoman dll. Makanya saya pengen ke Turki.	[]	Syiah	None
62225920	1212220317403250690	2020-01-01 12:44:17 WIB	2020-01-01	12:44:17	2545014338	alrasyidstore	abdulkarim adzdzahabi	@dyla_se @SKlaMud @yusuf_dumdum Jebule si syiah @yusuf_dumdum kakeane tenan.	[]	Syiah	None
82827521	1212666260372127745	2020-01-02 20:58:51 WIB	2020-01-02	20:58:51	428899478	burhanudin1994	Burhanudin	@Restichayah @HusinShihab Yg dipikirin si syiah @HusinShihab cmn selangkangan makannya gk bs membadakan wilayah dki dan banten	[]	Syiah	None

Gambar 3. 12 Add Column 'Topic Syiah' and 'Hate Speech'

Pada Gambar 3.11 menunjukkan adanya pembersihan *data* dan penambahan kolom baru yaitu '*Topic'* dan '*Hate Speech'*. Pada kolom '*Topic'* berisi *keyword* yang menjadi parameter pada proses *scraping* sebelumnya yaitu 'Cina', sedangkan pada Gambar 3.12 berisi *keyword topic* 'Syiah' dan Gambar 3.13 menunjukkan adanya penambahan kolom *topic* dengan kata kunci 'Ahmadiyah'. Lalu pada kolom '*Hate Speech'* merupakan kolom yang merepresentasikan label atau anotasi dari *Tweet* yang ada, apakah *Tweet* mengandung ujaran kebencian atau tidak. Jika *Tweet* mengandung ujaran kebencian maka diberikan *value* 1, jika tidak mengandung ujaran kebencian diberi *value* 0.

```
In [6]: pd.set_option('display.max_colwidth', None)
        df['topic']= 'Ahmadiyah'
        df['hatespeech']= None
        df
```

Out[6]:

onversation_id	created_at	date	time	user_id	username	name	tweet	hashtags	topic	hatespeech
224369397760	2020-01-06 21:29:19 WIB	2020-01-06	21:29:19	517170383	rahmayusman	Rahmah	Setauku, Syiah yang beneran itu baik-baik aja, gak sesat, gak seperti yang diberitakan banyak orang Sama kayak Ahmadiyah, mereka aslinya baik dan gak melenceng ajarannya gak kaya yang diberitakan.	[]	Ahmadiyah	None
610562736128	2020-01-07 11:42:15 WIB	2020-01-07	11:42:15	185402933	ordersofheavens	Mvly▲bvdi	Ada juga kemaren-2 yg masih musuhin ahmadiyah. Taunya lu dikasih nasib pergi ke Inggris dan ibadah Sholat dimesjid sana. mostly Ahmadiyah lah. Ada banyak Sunni jg yg membangga-2kan ilmuwan Islam Klasik. Mostly ilmuwan Syiah lah. Banyak ya cerita goblok goblok gobiok	[]	Ahmadiyah	None
588609269761	2020-01-17 10:45:59 WIB	2020-01-17	10:45:59	732263036	kata_dhoni	ali romdhoni	■ Muslim Television Ahmadiyah International #fkubjawatengah @ Kampus Mubarak https://t.co/mKB7mBytd3	['fkubjawatengah']	Ahmadiyah	None
							■ Pulng Masjid yang Dibakar Rombongan dari FKUB Jawa Tengah berpose di halaman			

Gambar 3. 13 Add Column 'Topic Ahmadiyah' and 'Hatespeech'

Out[4]:

e	name	place	...	reply_to	retweet_date	translate	trans_src	trans_dest	uncleaned_tweet	topic	hatespeech	wordscount	contains_link
s	Rach Alida Bahaweres	NaN	...	[]	NaN	NaN	NaN	NaN	@CommuterLine pagi ini dari Pondok Cina ke Ta...	Cina	0	14	0
ri	fachri52	NaN	...	[{'screen_name': 'liem_id', 'name': 'Mbah_Liem...	NaN	NaN	NaN	NaN	@liem_id Jd hongkong bkn cina ya.. br tau	Cina	1	8	0
a	dhik	NaN	...	[]	NaN	NaN	NaN	NaN	India, cina, medan, manado, ntt, depok, pwk, b...	Cina	1	16	1
a	Sabriel Rayyan	NaN	...	[{'screen_name': 'ACTforHumanity', 'name': 'ac...	NaN	NaN	NaN	NaN	@ACTforHumanity @Ajeng__Cute16 Kelurahan bidar...	Cina	0	24	1
s	Syahril Sidik	NaN	...	[{'screen_name': 'yarraPD', 'name': 'Septyara ...	NaN	NaN	NaN	NaN	@yarraPD Ya allah, pondasinya ga kokoh, soalny...	Cina	1	9	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
t	Gosip Garut	NaN	...	[]	NaN	NaN	NaN	NaN	Pelarangan Aktivitas dan Pembangunan Masjid Ah...	Ahmadiyah	0	18	1
n	Majlis Khuddam-ul-Ahmadiyyah Kebayoran	NaN	...	[]	NaN	NaN	NaN	NaN	Siaran Pers Khalifah Ahmadiyah Menyuarakan Hak...	Ahmadiyah	0	37	1
n	jaarrr	NaN	...	[]	NaN	NaN	NaN	NaN	Syiah bahasan lama, jauh sebelum Ahmadiyah ten...	Ahmadiyah	0	32	0
a	Tahun2021	NaN	...	[]	NaN	NaN	NaN	NaN	Kadang mikir ngene, sekarang santai ketemu ora...	Ahmadiyah	1	20	0

Gambar 3. 14 Add Column 'wordscount' and 'contains_link'

Gambar 3.14 menunjukkan adanya penambahan kolom '*wordscount'* dan '*contains_link'* pada *dataframe.* Kolom *wordscount* berisi jumlah kata yang terdapat di dalam *tweet*, sedangkan kolom *contains_link* berisi *value* 1 atau 0, dimana 1 yang berarti *tweet* mengandung *link* dan 0 yang berarti *tweet* tidak mengandung *link.*

```
In [13]: url = "https://docs.google.com/spreadsheets/d/1mAOoZW_aPmaAu_gOrBGM8CubgQukqD7lCf3sYLKTpK0/export?format=csv&gid=0"
         annotate = pd.read_csv(url)
         keywords=annotate['hatespeech'].tolist()

In [14]: keywords = [x for x in keywords if str(x) != 'nan']

In [15]: keywords

Out[15]: ['Goblok',
          'Anjing',
          'Komunis',
          'Jorok',
          'Benci',
          'Antek',
          'Kafir',
          'Kapir',
          'Loser',
          'Jongos',
          'Biadab',
          'Babi',
          'Pukimak',
          'Cebong',
          'Pki',
          'Mampos',
          'Tolol',
          'Bego',
          'Bangsat',
          'Usir',
          'Sipit',
          'Pesek',
          'Bangke',
          'Kontol',
          'Pelacur',
          'Asuuuu',
          'Pantat',
          'Sesat',
          'Pengkhianat',
          'Penghianat',
          'Tai',
```

Gambar 3. 15 Keyword Hate Speech

Pada gambar 3.15 menunjukkan adanya proses penambahan *keyword hate speech* dimana sumber *keyword* berasal dari *google docs* dan di *integrate* dengan bahasa pemrograman Python. Berikut 63 *keywords* yang mengandung *Hate Speech* antara lain 'Goblok', 'Anjing', 'Komunis', 'Jorok', 'Benci', 'Antek', 'Kafir', 'Kapir', 'Loser', 'Jongos', 'Biadab', 'Babi', 'Pukimak', 'Cebong', 'Pki', 'Mampos', 'Tolol', 'Bego', 'Bangsat', 'Usir', 'Sipit', 'Pesek', 'Bangke', 'Pelacur', 'Asuuuu', 'Sesat', 'Pengkhianat', 'Penghianat', 'Tai', 'Kacung', 'Anjeng', 'Babu', 'Neraka', 'Bacot', 'Kroyok', 'Penjilat', 'Kampret', 'Bodoh', 'Dasar', 'Fuck', 'Fake', 'Hell', 'Stupid', 'Terrorist', 'Ancaman', 'Bodo', 'Hancurkan', 'Hancur', 'Laknatullah', 'Fuckyou', 'Kubur', 'Fitnah', 'Dajjal', 'Ular', 'Kejahatan', 'Kesesatan', 'Bitch', 'Pedofil', 'Musuh', 'Setan', 'Fucking', 'Shaiton', 'Anjir', 'Perang'. *Tweet* yang mengandung *keyword* tersebut akan dianotasi menjadi *value* 1 pada kolom *hatespeech*. Sedangkan, *Tweet* yang tidak mengandung *keyword* akan dianotasi sebagai *value* 0 atau tidak mengandung ujaran kebencian pada kolom *hatespeech*. Hasil dari *labelling Hate Speech* dapat dilihat pada Gambar 3.16.

```
In [13]: df1 = df[(df.hatespeech == 1)]

In [14]: df1

Out[14]:
```

name	place	...	user_rt	retweet_id	reply_to	retweet_date	translate	trans_src	trans_dest	uncleaned_tweet	topic	hatespeech
Al Ghuyubi	NaN	...	NaN	NaN	[{'screen_name': 'daniarti03', 'name': 'water_ecosystem', 'id': '210732262'}, {'screen_name': 'Hanifah933', 'name': '[illegible] Hanifah Andini [illegible]', 'id': '1128507001346990080'}, {'screen_name': 'aniesbaswedan', 'name': 'Anies Baswedan', 'id': '110312278'}]	NaN	NaN	NaN	NaN	@daniarti03 @Hanifah933 @aniesbaswedan Rusia perkembangan muslim pesat cuyyy Komunis sejati tuh cina komunis@ Lu ga suka Islam? Fakta Islam turun di Arab, mau merubah fakta? Atau mau kaya cina komunis yg mau rubah AlQuran dan Injil sesuain paham komunis? Halu!!!	Cina	1.0
Al Ghuyubi	NaN	...	NaN	NaN	[{'screen_name': 'daniarti03', 'name': 'water_ecosystem', 'id': '210732262'}, {'screen_name': 'Hanifah933', 'name': '[illegible] Hanifah Andini [illegible]', 'id': '1128507001346990080'}, {'screen_name': 'aniesbaswedan', 'name': 'Anies Baswedan', 'id': '110312278'}]	NaN	NaN	NaN	NaN	@daniarti03 @Hanifah933 @aniesbaswedan Hehe mana ada khilafah produksi teroris.. Banyakin baca tong... Kekerasan / teror / ada dimana?.. Budha Nyanmar terhadap rohingya Hindu India terhadap Khasmir Cina Komunis terhadap Uyghur Zionis terhadap palestina ISIS (bentukan USA) terhadap muslim timut tengah Lu sehat?	Cina	1.0
AaronFrans	NaN	...	NaN	NaN	[]	NaN	NaN	NaN	NaN	@Regyalfriano Bacot lu cina.	Cina	1.0

Gambar 3. 16 Tweet Cina yang mengandung Hate Speech

```
In [3]: df1 = pd.read_csv('SyiahAnnotate.csv')

In [6]: df1 = df1[(df1.hatespeech == 1)]
        df1
```

user_id	username	name	place	...	user_rt	retweet_id	reply_to	retweet_date	translate	trans_src	trans_dest	uncleaned_tweet	topic	hatespeech
448287	adepedia	prajna	NaN	...	NaN	NaN	[]	NaN	NaN	NaN	NaN	Kemudian ada survei persepsi masyarakat Indone...	Syiah	1.0
715302	imanfs165	Firman_TheLawyer	NaN	...	NaN	NaN	[]	NaN	NaN	NaN	NaN	Sy bukan pendukung Iran, & sy pun bukan be...	Syiah	1.0
737996	fuadjamil	fuad jamil	NaN	...	NaN	NaN	[{'screen_name': 'discernia', 'name': 'DISCERNIA', ...	NaN	NaN	NaN	NaN	@discernia @MuhammadiyinGL Iran, rezim Syria, ...	Syiah	1.0
737996	fuadjamil	fuad jamil	NaN	...	NaN	NaN	[{'screen_name': 'ainuroh_mah', 'name': 'Ainu...	NaN	NaN	NaN	NaN	@ainuroh_mah @NUgarisucu @KatolikG @Muhammad...	Syiah	1.0

Gambar 3. 17 Tweet Syiah yang mengandung Hate Speech

```
In [4]: df1 = df1[(df1.hatespeech == 1)]
        df1

Out[4]:
```

name	name	place	...	user_rt	retweet_id	reply_to	retweet_date	translate	trans_src	trans_dest	uncleaned_tweet	topic	hatespeech
nutaqi	AS MUTAQI #S4LAM	NaN	...	NaN	NaN	[]	NaN	NaN	NaN	NaN	Satu Harapan: Observasi FKUB Jateng: Ahmadiyah...	Ahmadiyah	1.0
anstak	RH. Juntak M15	NaN	...	NaN	NaN	[{'screen_name': 'syarifjogja', 'name': 'syari...	NaN	NaN	NaN	NaN	@syarifjogja @B4ngTovyib Urus aja dulu perdama...	Ahmadiyah	1.0
nutaqi	AS MUTAQI #S4LAM	NaN	...	NaN	NaN	[]	NaN	NaN	NaN	NaN	Satu Harapan: Observasi FKUB Jateng: Ahmadiyah...	Ahmadiyah	1.0
anstak	RH. Juntak M15	NaN	...	NaN	NaN	[{'screen_name': 'syarifjogja', 'name': 'syari...	NaN	NaN	NaN	NaN	@syarifjogja @B4ngTovyib Urus aja dulu perdama...	Ahmadiyah	1.0
nchan	S O N Y	NaN	...	NaN	NaN	[]	NaN	NaN	NaN	NaN	Kalau ini anda akan berurusan dengan NU Sikap...	Ahmadiyah	1.0
n3482	Fireman348	NaN	...	NaN	NaN	[]	NaN	NaN	NaN	NaN	Toleransi Ke Ahmadiyah. Ke Syiah..ke Non..Gili...	Ahmadiyah	1.0
erkah	hasnaBERKAH	NaN	...	NaN	NaN	[{'screen_name': 'DimitriYusuf', 'name': 'Dimi...	NaN	NaN	NaN	NaN	@itacimedara_56 @DimitriYusuf @kumparan Gak ya...	Ahmadiyah	1.0
ribadi	holy money	NaN	...	NaN	NaN	[{'screen_name': 'simple_heart88', 'name': 'Ka...	NaN	NaN	NaN	NaN	@simple_heart88 Pada 2011 massa FPI membantai...	Ahmadiyah	1.0
ribadi	holy money	NaN	...	NaN	NaN	[{'screen_name': 'daveastrawinata', 'name': 'D...	NaN	NaN	NaN	NaN	@daveastrawinata @alHafizhi @Dandhy_Laksono	Ahmadiyah	1.0

Gambar 3. 18 Tweet Ahmadiyah yang mengandung Hate Speech

Hasil dari proses anotasi menggunakan *keyword* sebelumnya dapat dilihat pada Gambar 3.16 untuk *Tweet* Cina yang mengandung *Hate Speech*, Gambar 3.17 untuk *Tweet* Syiah yang mengandung *Hate Speech,* dan Gambar 3.18 menunjukkan *Tweet* Ahmadiyah yang mengandung *Hate Speech*. Pada hasil anotasi akan memberikan nilai 1 kepada *Tweet* yang mengandung ujaran kebencian dan nilai 0 kepada *Tweet* yang tidak mengandung ujaran kebencian.

```
In [103]: # Create a function to clean the tweets
def cleanTxt(text):
    text = re.sub(r'@[A-Za-z0-9]+', '', text) # Remove @mentions
    text = re.sub(r'[0-9]+', '', text) # Remove number
    text = re.sub(r'#','',text) # Remove #hashtag symbol
    text = re.sub(r'[^\w]', ' ', text)# Remove symbols
    text = re.sub(r'_', ' ', text)# Remove underscore
    text = re.sub(r'[\n]+', '', text) # Remove new line
    text = re.sub(r':', '', text) # Remove : symbols
    text = re.sub(r'RT[\s]+', '', text) # Remove RT word
    re.sub(r'^https?:\/\/.*[\r\n]*', '', text, flags=re.MULTILINE) # Removing https hyperlink
    re.sub(r'^http?:\/\/.*[\r\n]*', '', text, flags=re.MULTILINE) # Removing http hyperlink

    return text

# Clean the tweets
df['tweet'] = df['tweet'].apply(cleanTxt)

# Show the cleaned tweets
df
```

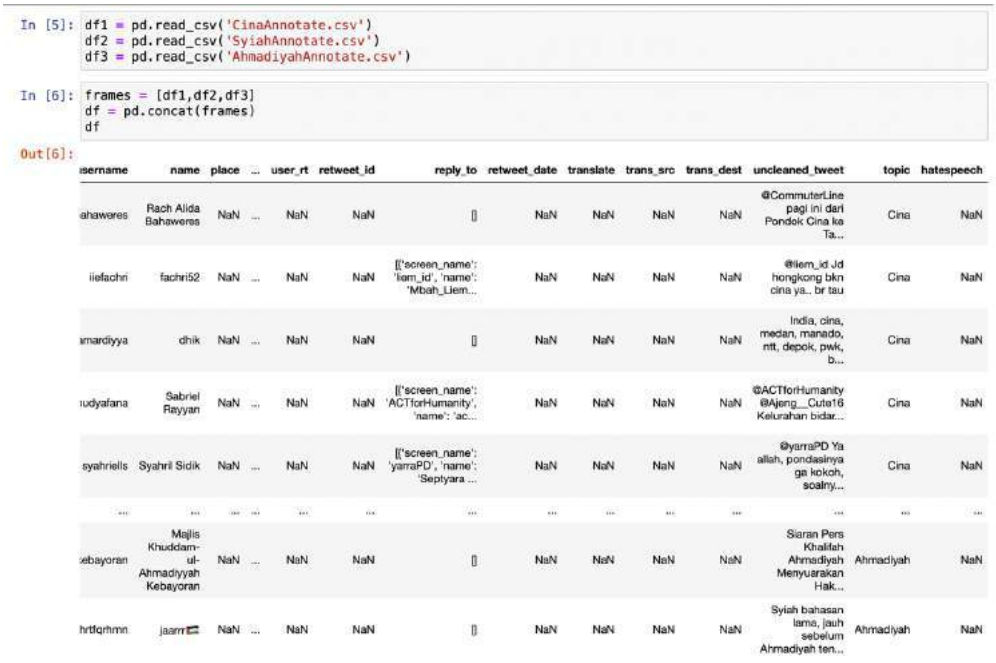
Out[103]:

	id	conversation_id	created_at	date	time	user_id	username	name	tweet	hashtags	tc
0	1212475813926322176	1212475813926322176	2020-01-02 03:48:55 WIB	2020-01-02	03:48:55	146038671	lidbahaweres	Rach Alida Bahaweres	pagi ini dari Pondok Cina ke Tanah Abang jam pagi apakah sudah normal Tq	[]	C
1	1212468631025111040	1212178324602011648	2020-01-02 03:00:30 WIB	2020-01-02	03:00:30	1243362931	ilefachri	fachri52	id Jd hongkong bkn cina ya br tau	[]	C

Gambar 3. 19 Regular Expression pada Dataframe

Pada gambar 3.19 menunjukkan adanya proses *data* manipulasi dengan menggunakan *regular expression* atau regex. Dalam proses tersebut menghapus beberapa *text* dari *Tweet* yang telah didapatkan dari proses *scraping*, contohnya menghapus *mentions*, angka, *hastags*, *symbol*, dan *underscore*. Penggunaan *syntax regular expression* dapat dilihat pada beberapa *dictionary* implementasi *regex*. Fungsi *regular expression* yang telah dibuat diberi nama *function cleanTxt. Function* tersebut diimplementasi ke dalam kolom *dataframe* yang berisi *Tweet* hasil *scraping*. Implementasi *regular expression* dilakukan ke semua *output scraping* dari topik Cina, Syiah, dan Ahmadiyah.

Pada Gambar 3.20 menunjukkan adanya penggabungan *data* dari semua topik *tweet* yang telah di *scraping*. Hasil dari *concat* tersebut berupa *file* csv dan siap diberikan kepada *Data Scientist* untuk diolah menjadi model *machine learning*.



```
In [5]: df1 = pd.read_csv('CinaAnnotate.csv')
        df2 = pd.read_csv('SyiahAnnotate.csv')
        df3 = pd.read_csv('AhmadiyahAnnotate.csv')

In [6]: frames = [df1,df2,df3]
        df = pd.concat(frames)
        df

Out[6]:
```

username	name	place	...	user_rt	retweet_id	reply_to	retweet_date	translate	trans_src	trans_dest	uncleaned_tweet	topic	hatespeech
ahaweres	Rach Alida Bahaweres	NaN	...	NaN	NaN	[]	NaN	NaN	NaN	NaN	@CommuterLine pagi ini dari Pondok Cina ke Ta...	Cina	NaN
iiefachri	fachri52	NaN	...	NaN	NaN	[{'screen_name': 'liem_id', 'name': 'Mbah_Liem...	NaN	NaN	NaN	NaN	@liem_id Jd hongkong bkn cina ya.. br tau	Cina	NaN
mardiyya	dhik	NaN	...	NaN	NaN	[]	NaN	NaN	NaN	NaN	India, cina, medan, manado, ntt, depok, pwk, b...	Cina	NaN
udyafana	Sabriel Rayyan	NaN	...	NaN	NaN	[{'screen_name': 'ACTforHumanity', 'name': 'ac...	NaN	NaN	NaN	NaN	@ACTforHumanity @Ajeng__Cute16 Kelurahan bidar...	Cina	NaN
syahriells	Syahril Sidik	NaN	...	NaN	NaN	[{'screen_name': 'yarraPD', 'name': 'Septyara ...	NaN	NaN	NaN	NaN	@yarraPD Ya allah, pondasinya ga kokoh, soalny...	Cina	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...
ebayoran	Majlis Khuddam-ul-Ahmadiyyah Kebayoran	NaN	...	NaN	NaN	[]	NaN	NaN	NaN	NaN	Siaran Pers Khalifah Ahmadiyah Menyuarakan Hak...	Ahmadiyah	NaN
hrtfqrhmn	jaarrr	NaN	...	NaN	NaN	[]	NaN	NaN	NaN	NaN	Syiah bahasan lama, jauh sebelum Ahmadiyah ten...	Ahmadiyah	NaN

Gambar 3. 20 Concat Data from All Topic

3.3.3 Hasil Analisis Sentimen

Pada hasil analisis sentiment terhadap 2 kategori *dashboard* yaitu *General Trends* dan *Content Analysis*. Pada *dashboard General Trends* memberikan visualisasi tren ujaran kebencian melalui *Twitter* di Indonesia termasuk tren mengenai *volume Tweet* ujaran kebencian berdasarkan waktu (per jam, per minggu, dan per bulan) serta tingkat keterlibatan terhadap ujaran kebencian dan rasio konten ujaran kebencian kepada kelompok ketiga kelompok Ahmadiyah, Syiah, dan Cina. *Data* yang ditampilkan dapat disaring berdasarkan rentang waktu dan kategori topik. Sedangkan pada *dashboard content analysis* menganalisis konten *tweet* yang mengandung ujaran kebencian, termasuk informasi mengenai *ratio tweet* yang menggunakan video, *link*, dan *hashtags* didalam *tweet*. Rincian *dashboard* yang terbagi dalam CSIS *National Hate Speech Dashboard* antara lain:

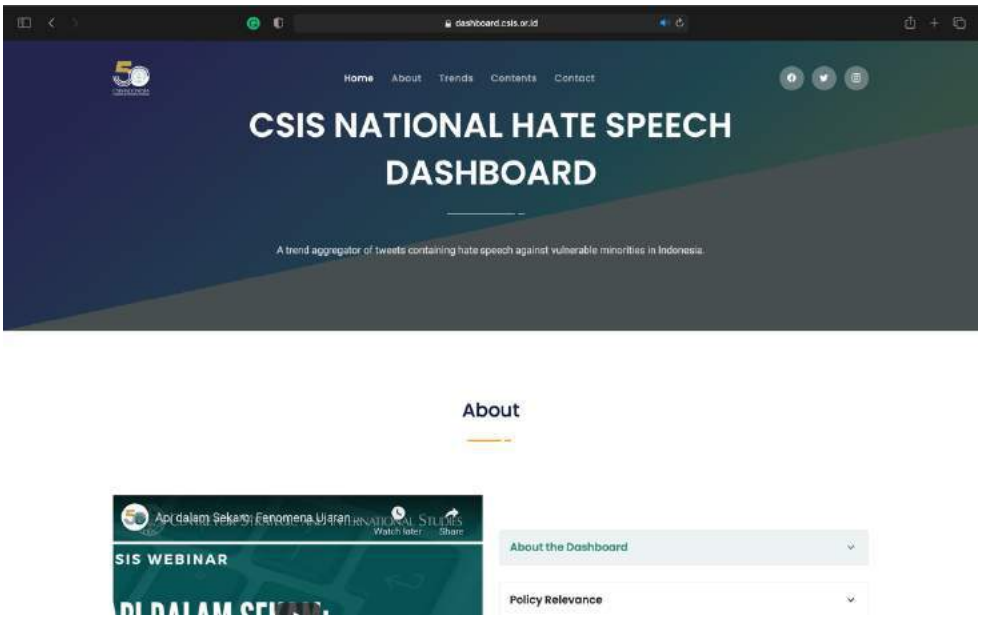
1) *General Trends* terdiri dari *dashboard*:
    - *Ratio of Tweets Targeting Vulnerable Communities: Ahmadiyyah, Shi'a, and Chinese Indonesians*

- *Monthly Number of Tweets Targeting Vulnerable Communities: Ahmadiyyah, Shi'a, and Chinese Indonesians*
- *Engagement Ratio of Tweets Containing Hate Speech Against Vulnerable Communities*
- *Monthly Number of Tweets Containing Hate Speech Against Vulnerable Communities*

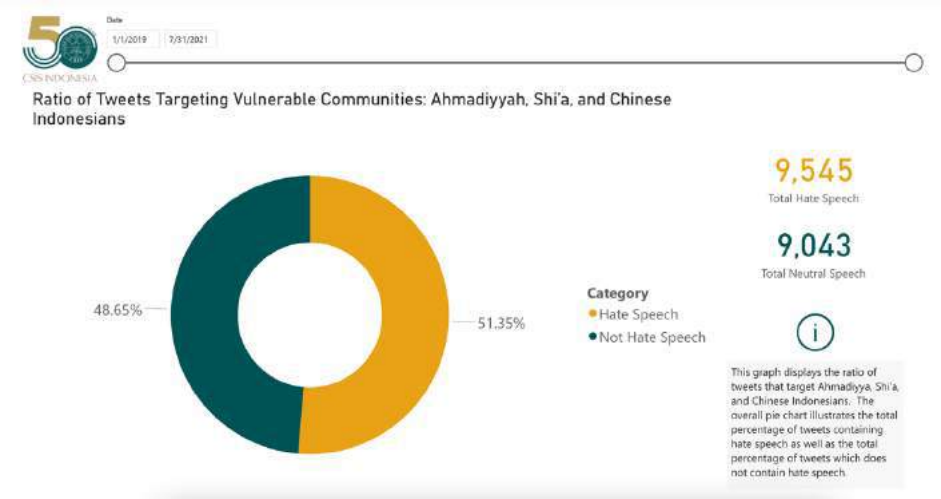
2) *Content Analysis* terdiri dari *dashboard*:

- *Ratio of Hate Speech Tweets Targeting Vulnerable Communities Which Contains Links and Videos*
- *Hashtags Used in Tweets Containing Hate Speech Against Vulnerable Communities*
- *Issues Linked to Tweets Containing Hate Speech Against Chinese Indonesians*

Gambar 3.21 menunjukkan *website* CSIS Indonesia yang menampilkan CSIS *National Hate Speech Dashboard*.

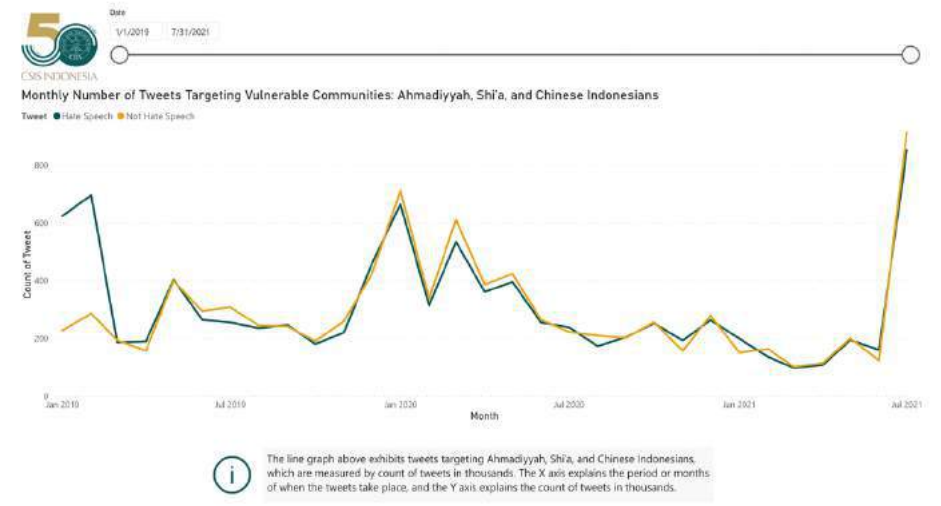


Gambar 3. 21 Website CSIS National Hate Speech Dashboard



Gambar 3. 22 Ratio of Tweets Dashboard

Pada Gambar 3.22 terdapat *dashboard "Ratio of Tweets Targeting Vulnerable Communities : Ahmadiyah, Syiah, and Chinese Indonesians"* yang menunjukkan *ratio tweet* dengan periode waktu dari 1 Januari 2019 hingga 31 Juli 2021 terdapat 9.545 *hate-speech* kepada komunitas Ahmadiyah, Syiah, dan Cina di Indonesia dan 9.043 *tweet* netral atau bukan *hate-speech*. Berdasarkan visualisasi *data* tersebut dapat dilihat bahwa lebih dari 50% *tweet* mengandung ujaran kebencian terhadap kelompok Ahmadiyah, Syiah, dan Cina di Indonesia.



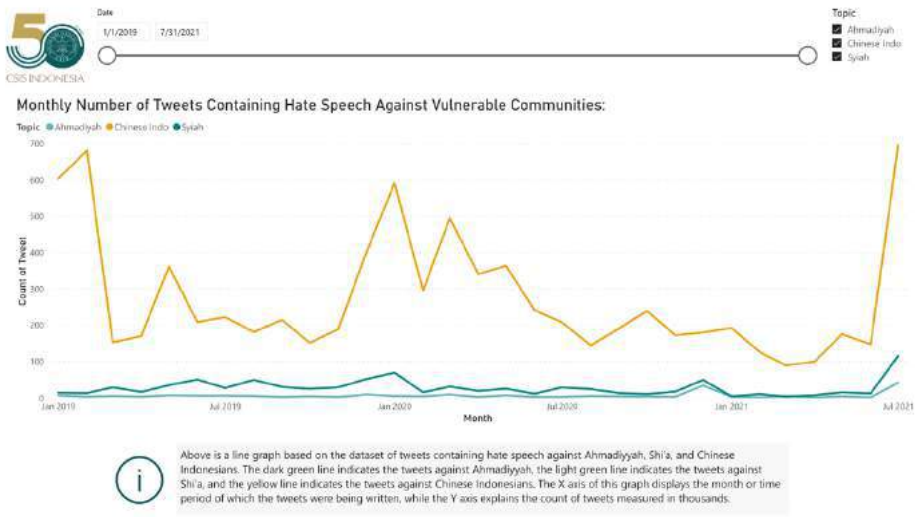
Gambar 3. 23 Monthly Number of Tweets Graph

Pada Gambar 3.23 terdapat tren perbulan total *hate-speech* di Indonesia terhadap kelompok Ahmadiyah, Syiah, dan Cina, apakah terjadi

peningkatan atau penurunan setiap bulannya, dimana total angka terendah *tweet* yang mengandung *hate-speech* adalah pada bulan Maret 2021 dengan jumlah sebesar 97 *tweet hate-speech*, sedangkan total angka tertinggi jumlah *tweet hate-speech* di Indonesia adalah pada bulan Juli 2021 dengan total 852 *tweet* yang mengandung ujaran kebencian.

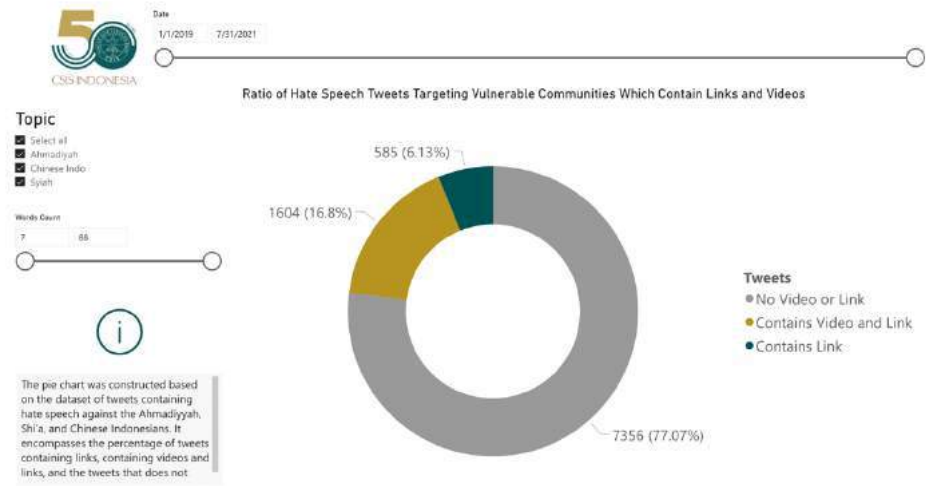
Gambar 3. 24 Engagement Ratio of Tweets Dashboard

Pada Gambar 3.24 adalah *dashboard* terkait *engagement ratio* pada *tweet* yang mengandung ujaran kebencian. Pada *dashboard engagement ratio* tersebut terdapat total *likes, replies, dan retweets.* Total angka pada *engagement ratio* dapat berubah sesuai topik masing-masing. Pada Gambar 3.21 menunjukkan adanya 34.438 total *likes* dari keseluruhan *tweet* yang mengandung ujaran kebencian. Selain itu terdapat 6.142 total *replies* dan 11.969 total *retweets* pada *tweet hate-speech* terhadap kelompok Ahmadiyah, Syiah, dan Cina di Indonesia.



Gambar 3. 25 Monthly Number of Tweets Dashboard

Pada Gambar 3.25 memberikan visualisasi terkait jumlah per bulan adanya *hate-speech* melalui media sosial Twitter terhadap kelompok Ahmadiyah, Syiah, dan Cina di Indonesia. Dalam *dashboard* tersebut menunjukkan bahwa kelompok Cina yang disimbolkan menggunakan garis kuning mendapatkan ujaran kebencian paling tinggi di antara kelompok lain dengan jumlah tertinggi sebanyak 695 *tweet hate-speech* pada bulan Juli 2021.



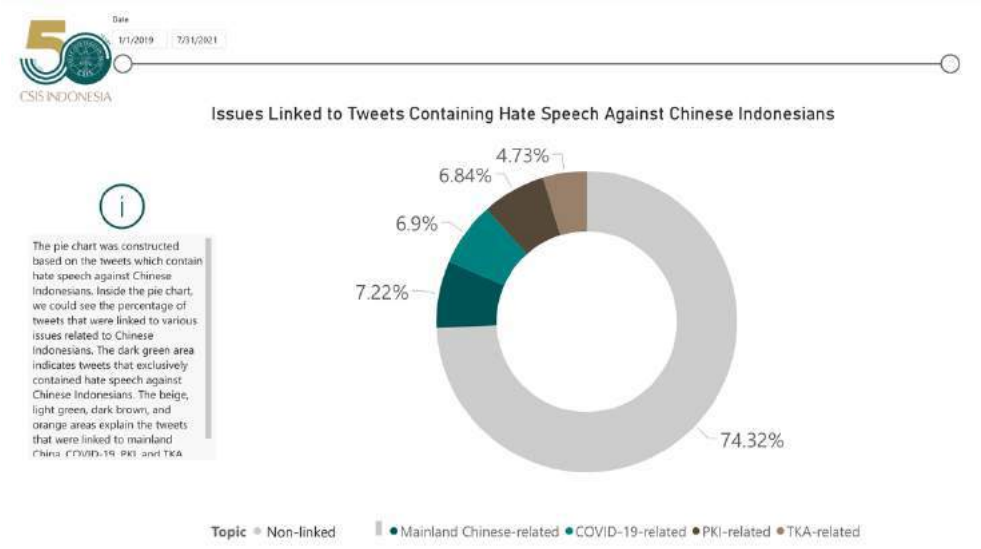
Gambar 3. 26 Ratio of Hate Speech Tweets Dashboard

Pada Gambar 3.26 mengandung informasi terkait *ratio* dari *tweet*, apakah *tweet hate-speech* mengandung video atau *link*, dan *hashtags*.

Berdasarkan visualisasi *data* tersebut, sebesar 77.07% *tweet* ujaran kebencian tidak mengandung video ataupun *link*, 16.8% mengandung video dan *link*, sedangkan *tweet hate-speech* yang hanya mengandung *link* hanya sebesar 6.13% dari total keseluruhan 9.545 *tweet hate-speech.*

Gambar 3. 27 Hashtags Visualization

Pada Gambar 3.27 menunjukkan visualisasi *word cloud* pada *hashtags* yang terdapat pada *tweet* ujaran kebencian, contohnya pada Gambar 3.24 merupakan hashtags yang paling banyak dipakai dalam ujaran kebencian terhadap kelompok Cina di Indonesia yaitu “coronavirus” , “tolakonebeltoneroad”, “tolakkolonialismekomunis”, “viruscorona”, “usir”, “sandiwaraanakmami” , dan “bangsatbangsa”.



Gambar 3. 28 Trend Tweet Dashboard

Gambar 3.28 menunjukkan *trend tweet* yang mengaitkan ujaran kebencian dengan beberapa kategori yaitu *Mainland Chinese*, COVID-19, PKI, dan TKA. Kategori tersebut dipilih oleh CSIS bertepatan dengan adanya kasus penyebaran COVID-19 di Indonesia. Pada *dashboard* tersebut menunjukkan sebesar 6.9% *tweet* berkaitan dengan topik COVID-19, 7.22% berkaitan dengan *Mainland Chinese*, 6.84% berkaitan dengan *tweet* terkait PKI, 4.73% *tweet* berkaitan dengan TKA dan sebesar 74.32% berkaitan dengan *Chinese Indonesian*. Daftar istilah yang dipakai pada *dashboard* diatas adalah sebagai berikut:

1) *Hate Speech*: *Tweet* yang menggunakan frasa yang melegitimasi tindakan permusuhan atau menganggap kualitas negatif terhadap identitas komunitas yang rentan yaitu, Ahmadiyah, Syiah, dan Tionghoa Indonesia.
2) *Engagement Ratio*: Jumlah *tweet* yang terlibat *likes, replied,* dan *retweet* dibandingkan dengan jumlah *tweet* yang tidak terlibat.
3) *Non-linked*: *Tweet* yang secara eksklusif berisi ujaran kebencian terhadap orang Tionghoa Indonesia.
4) *TKA-related*: *Tweet* yang berisi ujaran kebencian terhadap orang Tionghoa Indonesia saat mengaitkannya dengan pekerja asing Tionghoa.

5) *COVID-19-related*: *Tweet* yang berisi ujaran kebencian terhadap orang Tionghoa Indonesia saat mengaitkannya dengan Pandemi COVID-19 (termasuk vaksin).
6) *PKI-related*: *Tweet* yang berisi ujaran kebencian terhadap orang Tionghoa Indonesia yang terkait dengan Partai Komunis Indonesia (PKI) yang dilarang.
7) *Mainland Chinese-related*: *Tweet* yang berisi ujaran kebencian terhadap orang Tionghoa Indonesia saat menghubungkannya dengan Republik Rakyat Tiongkok.

Menurut CSIS Indonesia yang dikatakan dalam Webinar "Api dalam Sekam: Fenomena Ujaran Kebencian di Indonesia", Indonesia masih mengalami beragam *episode* ujaran kebencian terhadap kelompok minoritas. Melalui penelitian "Analisis Sentimen Ujaran Kebencian Terhadap Kelompok Ahmadiyah, Syiah, dan Cina di Indonesia" memberikan pola-pola umum dalam ujaran kebencian di Indonesia. Beberapa pemicu ujaran kebencian terhadap kelompok Ahmadiyah, Syiah, dan Cina di Indonesia adalah adanya kampanye terhadap Ahmadiyah pada tahun 2005 hingga 2011, kampanye terhadap Syiah pada tahun 2006 hingga 2012, dan yang terakhir adalah adanya kampanye terhadap Basuki Tjahaja Purnama pada tahun 2016 hingga 2017. Dalam semua kasus tersebut, minoritas rentan menjadi sasaran kekerasan, ancaman, dan berbagai peraturan yang akhirnya membatasi kemampuan mereka untuk menggunakan hak konstitusionalnya sebagai warga negara Indonesia. Pola-pola umum yang memicu terjadinya ujaran kebencian di Indonesia adalah :

1) Pola "Dinamika Kampanye" yang terdiri dari fase pemicu atau *trigger phase, escalation phase,* dan *normalization phase*. *Trigger phase* adalah fase dimana penghasut menggunakan sebuah kegiatan atau ucapan dari korban sebagai justifikasi untuk memulai ujaran kebencian. Pada umumnya, fase pemicu dampaknya *relative* masih dapat terkendali oleh polisi ataupun pemerintah daerah. Fase kedua adalah *escalation phase* dimana intensitas dari ujaran kebencian

meningkat bahkan sampai *level* nasional. Fase *escalation* dapat dikarakterisasi oleh 3 hal yaitu adanya transformasi, politisasi, dan mobilisasi. Fase yang terakhir adalah *normalization phase* dimana masyarakat menerima korban ujaran kebencian merasa berhak untuk diberikan ujaran kebencian.

2) Pola "Narasi dan Konten" yang *relative* sama yaitu narasi dengan adanya ancaman dan norma kelompok dominan. Dalam ketiga kampanye ujaran kebencian, penghasut menggambarkan minoritas sebagai pelaku melalui ucapan dan tindakannya yang mengancam norma-norma agama atau budaya kelompok dominan. Kebanyakan dari kampanye ujaran kebencian memanfaatkan hukum negara, khususnya UU PNPS No.1/1965 tentang penistaan agama dan membingkai minoritas sebagai pelanggar peraturan.
3) Pola "*Enabling Factors"* yaitu adanya pergeseran paradigma beragama dan berbudaya di masyarakat Indonesia, adanya insentif politik dan elektoral dalam pemilu tingkat lokal maupun nasional, dan tokoh-tokoh kelompok mayoritas merasa modal sosial-ekonomi mereka diancam oleh kelompok minoritas. Ujaran kebencian dimulai ketika penghasut merasa otoritas mereka di masyarakat lebih besar.

Melalui pola-pola tersebut, CSIS Indonesia memberikan rekomendasi kebijakan terhadap ujaran kebencian di Indonesia. Beberapa rekomendasi kebijakan tersebut antara lain:

1) Diperlukan sebuah sistem peringatan dini untuk ujaran kebencian di tingkat lokal. Berkaitan dengan hal tersebut "CSIS National Hate Speech Dashboard" berkontribusi dalam peringatan dini untuk memberikan kesadaran bagi masyarakat Indonesia terkait ujaran kebencian.
2) Memperbaiki peraturan yang ada agar tidak disalahgunakan oleh penghasut ujaran kebencian.
3) Memastikan politisi tidak menggunakan atau mendukung retorika ujaran kebencian di masa kampanye.

4) Memperluas paradigma beragama dan berbudaya yang inklusif-substantif dan bukan eksekutif-legal formalistik.

3.4 Kendala yang dihadapi

Kendala yang dihadapi pada saat melaksanakan kerja magang adalah:

1) Beberapa *requirement* yang berubah dan semakin kompleks dari sebelumnya. Hal ini membuat proses pengerjaan memakan waktu yang banyak dan membuat proses pengerjaan menjadi lebih sulit dari sebelumnya.
2) Kurangnya pelatihan atau pengenalan terkait *Data Engineering Environment* di Blankspace.

3.5 Solusi atas Kendala

Solusi atas kendala yang ada adalah:

1) Solusi terhadap *requirement* yang berubah-ubah adalah berdiskusi dengan pihak client, apa saja hal-hal penting yang perlu diperhatikan dalam proses penelitian, seperti tujuan penelitian dan fokus penelitian beserta variabel penting dalam penelitian ini.
2) Mencari tahu dan inisiatif bertanya jika kurang paham atau belum mengerti.