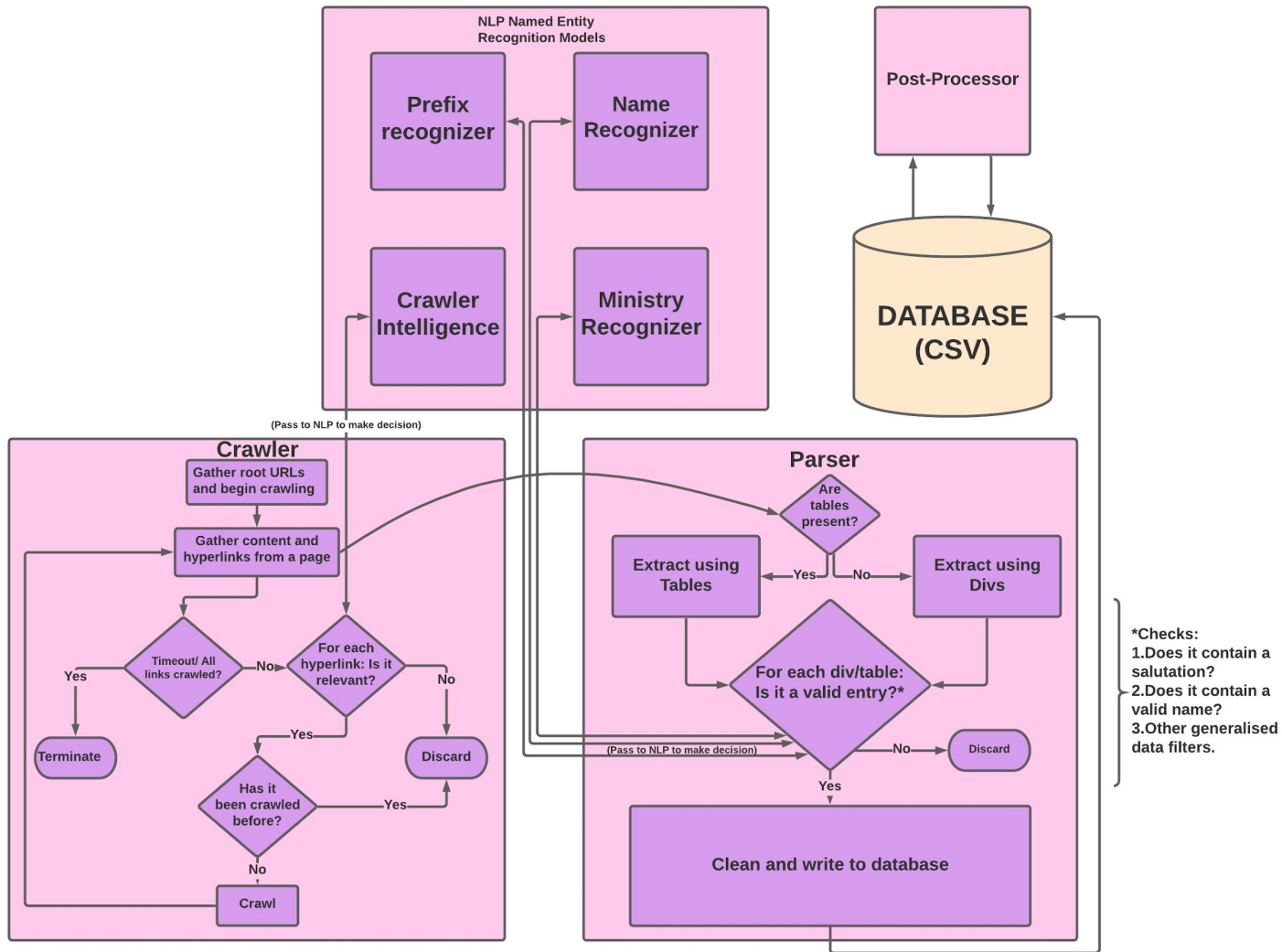
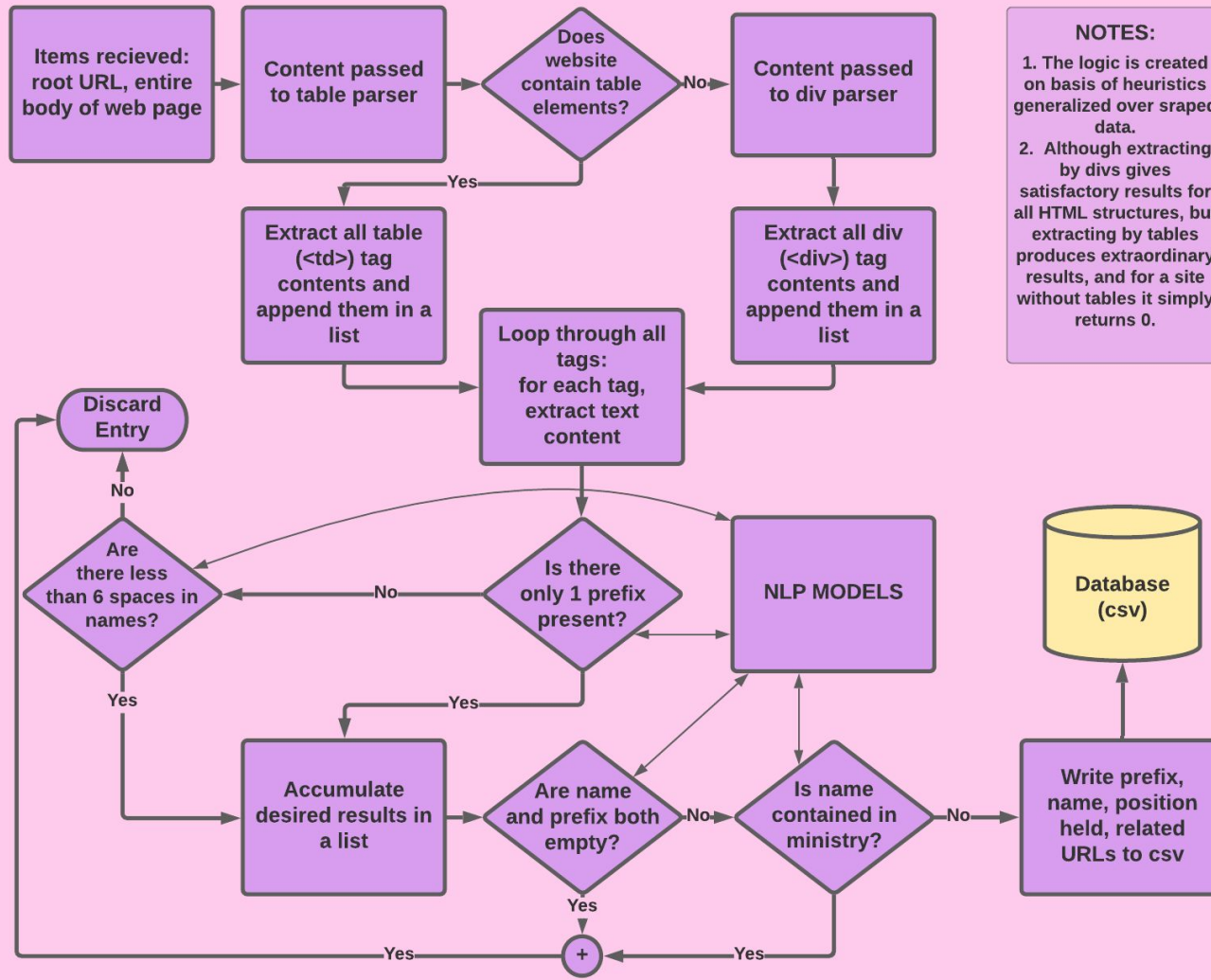


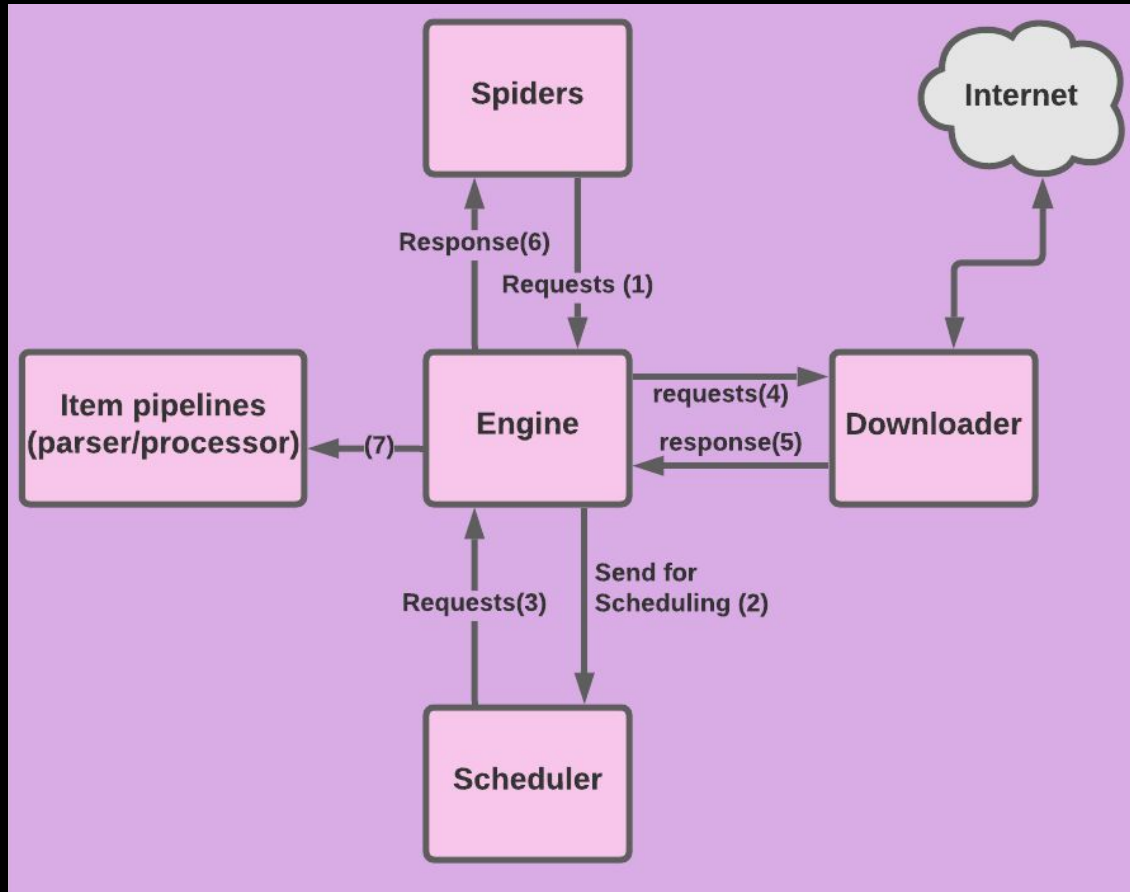
ThaparWings 2020

Team: Bicameral Minds

Team Members:
Avik Kuthiala
Naman Tuli







```

21 def pred(tag):
22     tag = tag.strip().lower()
23     stop_words = ["external website"]
24     for sw in stop_words:
25         if sw in tag:
26             return 0
27
28     if(tag.isnumeric()):
29         return 1
30     doc = nlpIntel(tag)
31
32     for ent in doc.ents:
33         if ent.text!=0:
34             #print(tag)
35             return 1
36     return 0

```

Crawler passes text content of <a> tag to this function to decide if it should be crawled or not.

1. If there is a strict domain that is not to be crawled, it is to be added in this list.
2. If the text is a number: crawl it.
Doing this ensures that multiple pages of a same domain are crawled (for eg. india.gov.in/lok-sabha/?page=2 must not be skipped).
3. If NLP model says crawl, we crawl.

```

90 def sc_divs(url, soup):
91     divs = soup.findAll("div")
92     result = []
93     repeat_check = []
94     for div in divs:
95         content = div.text.strip().replace("\n", "")
96         if content in repeat_check:
97             continue
98         repeat_check.append(content)
99
100         if has_name(content) == 0:
101             continue
102
103         tags = div.findAll("a", href=True)
104         links = [tag["href"] for tag in tags]
105         if links:
106             for i in range(len(links)):
107                 if "http" not in links[i]:
108                     links[i] = url + links[i]
109
110         result.append([content, links])
111     #print(len(repeat_check), repeat_check)
112     return result

```

1. Find all divs.
2. For each div, check for repeats.
3. If there is 0 result from has_name function, discard.
4. Find all associated links.
5. Return this 2D list.

```

def returner(string):
    doc = nlp_Name(string)
    name=""
    stop_words=['Shri','Smt','Smt.','Dr.','Dr','Mr','Mrs','Cabinet','Minister',
                'Contact','Facebook','Account'] #some very common stop-words
    for count,ent in enumerate(doc.ents):
        name+=ent.text+" "
        if count==0:
            break
    ret_name=""
    for words in name.split(" "):
        if words not in stop_words:
            ret_name+=words+" "
    doc =nlp_Min(string)
    ministry=""
    for ent in doc.ents:
        ministry+=ent.text+" "
    doc =nlp_Pref(string)
    prefix=""
    for ent in doc.ents:
        prefix+=ent.text+" "
    return prefix, ret_name, ministry

```

What this function does is:

Call the 3 NER* models extract Prefixes (Mr. Ms. etc), Names, position held (or ministry) and returns it to CSV writer.

*Named Entity Recognition.


```

def parse_soup(url, soup):
    if sc_table(url, soup) is not 0:
        content = sc_table(url, soup)
    else:
        content = sc_divs(url, soup)
    #return content
    write_obj = open("austria.csv", 'a+', newline='')
    csv_writer = writer(write_obj)
    for items in content:
        prefix,ret_name,ministry=returner(items[0])
        l1 = [prefix.strip(),ret_name.strip(),ministry.strip()]
        if l1[0] == '' and l1[1] == '': #no name no salutation //cleaner step
            continue
        if l1[1] in l1[2]: #name found in ministry //cleaner step
            continue
        l1.append(' '.join(items[1])) #urls
        csv_writer.writerow(l1)
    write_obj.close()

```

Search for tables, if yes, scrape tables, otherwise scrape divs.

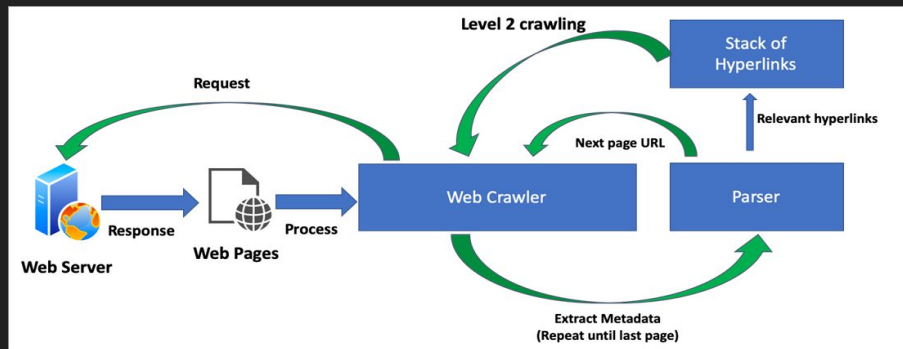
Do some more cleaning, write to CSV.

Crawler Intelligence

Since the domain of websites given to us is very large and most of the websites are deep and have only a fraction of the content is deemed scrapable, it becomes integral to devise a strategy that could solve this problem and make the efforts of the parser minimum. We have taken ideas from

<https://ieeexplore.ieee.org/document/7087203>

and have proposed a novel method consisting of a combined approach that solves the problem at hand.



The approaches have been enlisted as follows:

1. Lexicon based approach using RNN and LSTMs.
2. NER based approach implemented using Spacy for a custom trained dataset

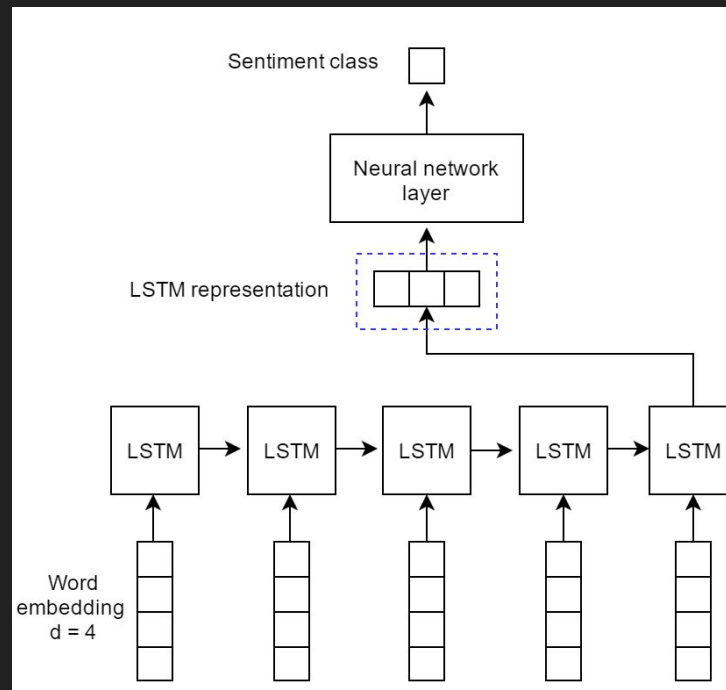
Lexicon Based Approach

We have manually curated a database that consists of 1500 lines of text of :

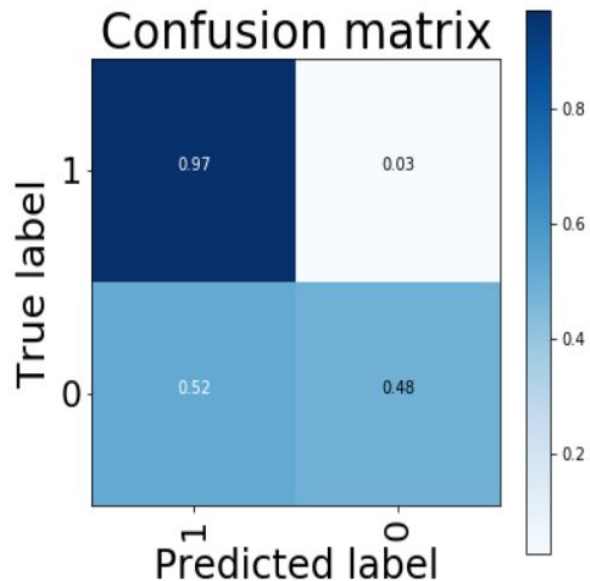
1. Words needed to be scraped (+ve)
2. Noise/Trash

```
['Directorate General of Commercial Intelligence and Statistics',  
'new-zealand-transport-agency',  
'new-zealand-defence-force',  
'city-rail-link-limited',  
'kiwirail']
```

```
['चंडीगढ़ में पशुपालन और मत्स्य पालन विभाग',  
'Information on Allotment of regular LPG distributorship to ESM/Widows',  
'campaign medal rolls',  
'Making payments through WorldPay']
```



Results on Sentiment Based Approach(3:1 Split)



ACCURACY: 0.7679192815792586

LOSS: 0.5050979286079699

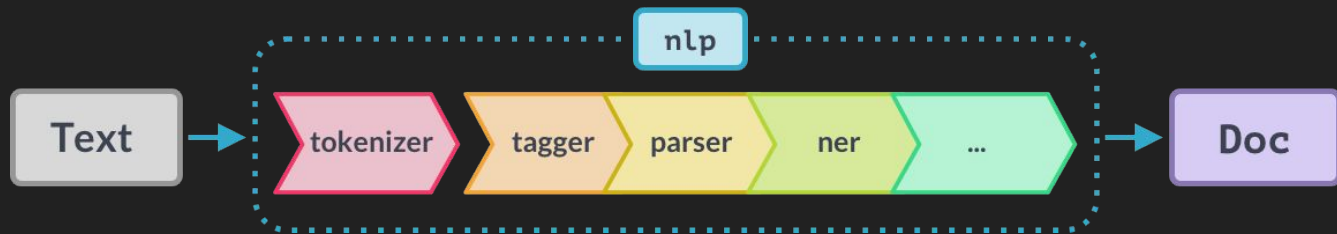
Drawbacks:

1. Overfitting due to lack of large enough training data
2. Some false values creeping in as true values
3. Mismatch in values
For eg.
Minister of Finance
Reports of Finance

Positives:

1. High TPR means that no true values are being neglected
2. If we ignore the false positives by setting a high probability threshold this could be used

Named Entity Recognition (NER)



NER is probably the only strategy that can be used to solve this problem of filtering the noise out of the data.

Among NER, we used both HuggingFace Transformers(BERT,ROBERTA,DISTILBERT) and spacy custom trained models for the problem.

But we discarded the former as the latter gave better results on the test data.

NER only extracts the useful information from the text and renders rest of the text unchanged.

RESULTS AND POSITIVES

```
TRAIN_DATA=['Ministry of Home Affairs',"Finance Ministry","Judiciary","Ministry of Women","Records","SignUp","Records","Documents"]  
hypothetical=["Ministry of Women and Finance Affairs","Ministry of ML","Intelligence Department"]  
language_barrier=[" Finance Dokumentar"]
```

```
for t in TRAIN_DATA:  
    print(output(t))
```

```
1  
1  
1  
1  
None  
None  
None  
None
```

```
▶ for t in hypothetical:  
    print(output(t))
```

```
↳ 1  
   1  
   1
```

```
▶ for t in language_barrier:  
    print(output(t))
```

```
↳ 1
```

NER Models for Entity Recognition

```
TEST=["Dr Husain Khan Contact Facebook Account Twitter Account Ministry of Agriculture & Farmers ",  
      "Ramaphosa, Matamela Cyril, Mr Presidency (The) ",  
      "His Highness Sheikh Mohammed Bin Rashid Al Maktoum Vice-President Prime Minister and Minister of Defence",  
      "Mr Luke Cooper Minister of Foreign Affairs",  
      "Minsiter of Health Mr Samir Handanovic"]
```

```
('Husain Khan ', 'Dr ', 'Ministry of Agriculture & Farmers ')  
( 'Ramaphosa Matamela Cyril ', 'Mr ', 'Presidency ' )  
( 'Mohammed Bin Rashid Al Maktoum ', 'His Highness Sheikh ', 'Vice-President Prime Minister and Minister of Defence ' )  
( 'Luke Cooper ', 'Mr ', 'Minister of Foreign Affairs ' )  
( 'Samir Handanovic ', 'Mr ', 'Minsiter of Health ' )
```


Consolidated Database India

76	Shri	Nitin	Ministry of Road Transport and Highways	Ministry of M	https://morth.nic.in/sites/default/files/Directory_21.01.2020.pdf	https://www.facebook.com/nitin
77	Shri	D.V.	Ministry of Chemicals and Fertilizers		https://chemicals.nic.in/sites/default/files/Minister_0_0.pdf	https://www.facebook.com/DVSBJP/
78	Smt.	Nirmala Sitharaman	Ministry of Finance	Ministry of Corporate Affairs	https://doe.gov.in/minister/meet-union-finance-minister	https://www.facebook.com/nirmala.sitha
79	Shri	Ramvilas Paswan	Ministry of Consumer Affairs, Food and Public Distribu		https://consumeraffairs.nic.in/about-us/whos-who	https://www.facebook.com/rvplojapa5/
80	Shri	Narendra Singh	Ministry of Agriculture & Farmers Welfare	Ministry of F	https://www.panchayat.gov.in/web/guest/who-s-is-who	https://www.facebook.com/narendrasing
81	Shri	Ravi Shankar Prasad	Ministry of Law and Justice	Ministry of Communication	http://meity.gov.in/content/cabinet-minister	https://www.facebook.com/RaviShankarPrasadOffici
82	Smt.	Harsimrat Kaur Badal	Ministry of Food		http://www.mofpi.nic.in/about-us/whos-who	https://www.facebook.com/Harsimratkaurbadal
83	Shri	Thaawar Chand Gehlot	Ministry of Social Justice and Empowerment		http://socialjustice.nic.in/UserView/index?mid=76721	https://www.facebook.com/thawarchand.ge
84	Dr.	Subrahmanyam Jaishankar	Ministry of External Affairs		https://meatel.nic.in/	https://twitter.com/drsjaishankar
85	Shri	Ramesh Pokhriyal 'I	Ministry of Education		https://mhrd.gov.in/whos-who	https://www.facebook.com/DrRPNishank/
86	Shri	Arjun Munda	Ministry of Tribal Affairs		https://tribal.nic.in/contactUs.aspx	https://www.facebook.com/arjunmunda/
87	Smt.	Smriti Zubin Irani	Ministry of Women and Child Development	Ministry of	http://texmin.nic.in/whos-who	https://www.facebook.com/Smriti.Irani.Official
88	Dr.	Harsh Vardhan	Ministry of Health and Family Welfare	Ministry of Scier	https://main.mohfw.gov.in/about-us/whos-who	https://www.facebook.com/drharshvardhanoffici
89	Shri	Prakash	Ministry of Environment, Forest and Climate Change	M	http://moef.gov.in/about-the-ministry/ministers/cabinet-minister-2/	https://www.facebook.com/F
90	Shri	Piyush	Ministry of Railways	Ministry of Commerce and Indust	http://www.indianrailways.gov.in/biodata_mr.html	https://www.facebook.com/PiyushGoyalOffici
91	Shri	Mukhtar Abbas	Ministry of Minority Affairs		http://www.minorityaffairs.gov.in/about-us/who-is--who	https://www.facebook.com/naqvimukht
92	Shri	Pralhad Joshi	Ministry of Parliamentary Affairs	Ministry of Coal Minis	https://mpa.gov.in/about-us/ministry-telephone-list	https://www.facebook.com/pralhadvjoshi/
93	Dr.	Mahendra Nath	Ministry of Skill Development and Entrepreneurship		https://www.msde.gov.in/contact-minister.html	https://www.facebook.com/drmnpandeymp
94	Shri	Giriraj Singh	Ministry of Animal Husbandry, Dairying and Fisheries		http://www.dahd.nic.in/about-us/whos-who	https://www.facebook.com/girirajsinghp
95	Shri	Gajendra Singh	Ministry of Jal Shakti		http://mowr.gov.in/minister/shri-gajendra-singh-shekhawat	https://www.facebook.com/mpjodhpu
96	Shri	Santosh	Ministry of Labour and Employment		http://labour.gov.in/whos-who	https://www.facebook.com/santosh.gangwar
97	Shri Shripad	Ayurveda, Yoga			https://main.ayush.gov.in/about-us/whos-who	https://www.facebook.com/shripadynaik/
98	Dr.	Jitendra Singh	Ministry of Development of North Eastern Region		http://mdoner.gov.in/content/minister-doner	https://www.facebook.com/drijitendras
99	Shri	Kiren	Ministry of Youth Affairs and Sports		https://yas.nic.in/contactus	https://www.facebook.com/KirenRijiju/
100	Shri	Raj Kumar Singh	Ministry of Power	Ministry of New and Renewable Ene	https://powermin.nic.in/en/content/contact-us-0	https://www.facebook.com/RajKumarSinghIndia
101	Shri	Hardeep Singh	Ministry of Housing and Urban Affairs	Ministry of Civil	http://mohua.gov.in/cms/telephone-directory.php	https://www.facebook.com/HardeepSPuri
102	Shri	Mansukh L. Mandaviya	Ministry of Shipping		http://sagarmala.gov.in/	https://www.facebook.com/mansukhmandviya
103	Dr.	Jitendra Singh			http://mdoner.gov.in/content/minister-doner	https://www.facebook.com/drijitendras
104	Shri	Kiren	Ministry of Minority Affairs		http://www.minorityaffairs.gov.in/about-us/who-is--who	https://www.facebook.com/KirenRijiju/
105	Shri	Raj Kumar Singh	Ministry of Skill Development and Entrepreneurship		https://www.msde.gov.in/contact-mos.html	https://www.facebook.com/RajKumarSinghIndia
106	Shri	Hardeep Singh	Ministry of Commerce and Industry		https://commerce.gov.in/OfficerContact.aspx	https://www.facebook.com/HardeepSPuri
107	Shri	Mansukh L. Mandaviya	Ministry of Chemicals and Fertilizers		https://chemicals.nic.in/sites/default/files/PHARMACEUTICALS.....docx%20-%20Shortcut.Ink_.pc	https://www.facebook.com/AshwiniKChouhev/
108	Shri	Faggansingh Kulaste	Ministry of Steel		https://steel.gov.in/telephone-directory	https://www.facebook.com/FSKulaste/
109	Shri	Ashwini Kumar Choube	Health and	Ministry of Health and Famil	https://main.mohfw.gov.in/about-us/whos-who	https://www.facebook.com/AshwiniKChouhev/

UAE

Sl. No.	Name	Position	Official Website	Social Media Links
39	Sheikh	Mohammed bin Rashid Al Maktoum	National Agenda	https://uaecabinet.ae/en/national-agenda
40	His Highness Sheikh	Home UAE Federal Supreme Council	UAE Federal Supreme Council	https://uaecabinet.ae/en https://uaecabinet.ae/en/uae https://uaecabinet.ae/en/uae
41	His Highness Sheikh	Home UAE Federal Supreme Council	UAE Federal Supreme Council	https://uaecabinet.ae/en https://uaecabinet.ae/en/uae https://uaecabinet.ae/en/uae
42	His Highness Sheikh	Home UAE Federal Supreme Council	UAE Federal Supreme Council	https://uaecabinet.ae/en https://uaecabinet.ae/en/uae https://uaecabinet.ae/en/uae
43	His Excellency	Sultan bin Saeed Al Badi	Cabinet Members Minister of Minister of Mini	https://twitter.com/alhammadihh https://twitter.com/nouraalkaabi https://www.instagram.com/alhammadihh
44	His Excellency	Sultan bin Saeed Al Badi	Minister of Minister of Minister of State for Dr	https://twitter.com/alhammadihh https://twitter.com/nouraalkaabi https://www.instagram.com/alhammadihh
45	His Excellency	Sultan bin Saeed Al Badi	Minister of Justice	
46	His Excellency	Hussain bin Ibrahim Al Hammadi	Minister of Education	https://twitter.com/alhammadihh
47	His Excellency	Mohammed bin Ahmad Al Bawardi	Minister of State for Defence Affairs	
48	Her Excellency	Mohammed Al Kaabi	Minister of Culture and Youth	https://twitter.com/nouraalkaabi https://www.instagram.com/nak/
49	Her Excellency	Sarah bint Yousif Al Amiri	Minister	https://twitter.com/SarahAmiri1 https://www.instagram.com/SarahAmiri1/
50	His Excellency	Omar bin Sultan Al Olama	Minister	https://twitter.com/omarsalolama/ https://www.instagram.com/omarsalolama/
51	HIS EXCELLENCY	AHMED ALI AL		
52	Her Excellency	Ohoud bint		https://twitter.com/ohoodalroumi https://www.instagram.com/ohoodalroumi/
53	Her Excellency	Suhail Faris Al Mazrui	Minister	https://twitter.com/shamma https://instagram.com/shamma
54	His Excellency	Zaki Anwar		https://twitter.com/Zakinus
55	Her Excellency	Mohammed Saeed Hareb Almheiri	Minister	https://twitter.com/mariamalmheiri https://www.instagram.com/mariamalmheiri/
56	His Excellency	Obaid bin Humaid Al Tayer	Minister of State for Financial Affairs	
57	His Excellency	Suhail bin Mohammed Faraj Faris Al Mazrouei	Minister of Energy and Infrastructure	https://twitter.com/hesuail https://www.instagram.com/shlbinfaraj/
58	His Excellency Sheikh	Nahayan Mabarak Al Nahayan	Minister of Tolerance and Coexistence	
59	His Excellency	Mohammed bin Abdullah Al Gergawi	Minister of Cabinet Affairs	
60	His Excellency	Ahmed Juma Al Zaabi	Minister	
61	His Excellency	Bin Mohamed Al Owais	Minister of Health and Prevention Minister of State for Federal National Council Affairs	
62	His Excellency Dr.	Anwar bin Mohammed Gargash	Minister of State for Foreign Affairs	https://twitter.com/anwargargash https://www.instagram.com/anwargargash/
63	His Excellency	Nasser bin	Minister of Human Resources and Emiratization	

SOUTH AFRICA

Dr	Ruth Segomotsi Mompoti District	Water and Sanitation welcomes court order issued	https://www.gov.za/media-statements/speeches/water-and-sanitation-welcomes-court-order-issued
Mr	Ramaphosa	Presidency	https://www.gov.za/about-government/leaders/about-government/contact-directory/matamela-cy
Mr	Mabuza	Presidency	https://www.gov.za/about-government/leaders/about-government/contact-directory/david-dabed
Ms	Didiza	Agriculture, Land Reform and Rural Development	https://www.gov.za/about-government/leaders/about-government/contact-directory/angela-thoko
Ms	Motshekga	Basic Education	https://www.gov.za/about-government/leaders/about-government/contact-directory/matsie-ange
Ms	Ndabeni-Abrahams	Communications and Digital Technologies	https://www.gov.za/about-government/leaders/about-government/contact-directory/tembisa-stel
Dr	Dlamini Zuma	Cooperative Governance and Traditional Affairs	https://www.gov.za/about-government/leaders/about-government/contact-directory/nkosazana-c
Ms	Mapisa-Nqakula	Defence and Military Veterans	https://www.gov.za/about-government/leaders/about-government/contact-directory/nosiviwe-no
Mr	Thembelani Thulas	Employment and Labour	https://www.gov.za/about-government/leaders/about-government/contact-directory/thembelani-t
Ms	Creecy	Environment, Forestry and Fisheries	https://www.gov.za/about-government/leaders/about-government/contact-directory/environment
Mr	Mboweni	Finance	https://www.gov.za/about-government/leaders/about-government/contact-directory/finance-mini
Dr]	Health	https://www.gov.za/about-government/leaders/about-government/contact-directory/cooperative
Dr	Nzimande	Higher Education, Science and Technology	https://www.gov.za/about-government/leaders/about-government/contact-directory/bonginkosi-e
Dr	Pakishe Aaron	Home Affairs	https://www.gov.za/about-government/leaders/about-government/contact-directory/pakishe-aard
Ms	Grace Naledi Mandisa	International Relations and Cooperation	https://www.gov.za/about-government/leaders/about-government/contact-directory/grace-naledi
Mr	Lamola	Justice and Correctional Services	https://www.gov.za/about-government/leaders/about-government/contact-directory/justice-and-c
Mr	Samson Gwede	Mineral Resources and Energy	https://www.gov.za/about-government/leaders/about-government/contact-directory/mineral-reso
Gen	Cele	Police	https://www.gov.za/about-government/leaders/about-government/contact-directory/bheki-cele-g
Mr	Mthembu, Jackson	Presidency	https://www.gov.za/about-government/leaders/about-government/contact-directory/presidency/j
Ms	Nkoana-Mashabane	Presidency for Women, Youth and Persons with Disabilities	https://www.gov.za/about-government/leaders/about-government/contact-directory/maite-nkoan
Mr	Gordhan	Public Enterprises	https://www.gov.za/about-government/leaders/about-government/contact-directory/pravin-jamna
Mr	Mchunu	Public Service and Administration	https://www.gov.za/about-government/leaders/about-government/contact-directory/senzo-mchu
Ms	de Lille	Public Works and Infrastructure	https://www.gov.za/about-government/leaders/about-government/contact-directory/public-works
Ms	Ntshavheni	Small Business Development	https://www.gov.za/about-government/leaders/about-government/contact-directory/small-busine
Ms]	Social Development	https://www.gov.za/about-government/leaders/about-government/contact-directory/lindiwe-zulu
Mr	Nkosinathi Emmanuel ¹	Sports, Arts and Culture	https://www.gov.za/about-government/leaders/about-government/contact-directory/nkosinathi-e
Ms	Dlodlo	State Security	https://www.gov.za/about-government/leaders/about-government/contact-directory/ayanda-dlod
Ms	Kubayi-Ngubane	Tourism	https://www.gov.za/about-government/leaders/about-government/contact-directory/mmamoloko
Mr	Patel, Ebrahim	Trade and Industry	https://www.gov.za/about-government/leaders/about-government/contact-directory/ebrahim-pat
Mr	Mbalula	Transport	https://www.gov.za/about-government/leaders/about-government/contact-directory/fikile-april-m
Mr	Skwatsha	Agriculture, Rural Development and Land Reform	https://www.gov.za/about-government/leaders/about-government/contact-directory/mcebisi-skwa
Mr	Agriculture	Agriculture, Rural Development and Land Reform	https://www.gov.za/about-government/leaders/about-government/contact-directory/agriculture-l
Ms	Mhaule	Basic Education	https://www.gov.za/about-government/leaders/about-government/contact-directory/international

CANADA

Dr.	Chris Alexander	Minister of Citizenship and Immigration	Minister of Citizenship and Immigration	https://www.canada.ca/en/immigration-refugees-citizenship/news/archives/speeches/2015/05/20150514-citizenship-immigration-minister-speech.html							
Dr.	Chris Alexander	Minister of Citizenship and Immigration	Minister of Citizenship and Immigration	https://www.canada.ca/en/immigration-refugees-citizenship/news/archives/speeches/2015/05/20150514-citizenship-immigration-minister-speech.html							
Dr.	Oyedeji Ayonrinde	Associate	Department of Psychiatry	Queen's University							
Dr.	Elaine Hyshka	Assistant	Health Policy and Management	School of Public Health, University of Alberta	A health services expert who focuses on advancing						
Dr.	Didier Jutras-Aswad	Psychiatrist									
Dr.	Amy Porath	Director									
Dr.	Maude St-Onge	Medical									
Dr.	Phil Tibbo	Professor	Department of Psychiatry	Dalhousie University	A						
Dr.	Mark Ware	Associate	Family Medicine and Anesthesia	McGill University	A						
Mr.	James Wigmore	Forensic Toxicologist	Mr. James Wigmore	Forensic Toxicologist							