VISIT THE DOME – WEEK 3

Thang Nguyen

Vamsi Vaddi

Ruta Bakhda

TASK

- Task 1: Identify political party as Government or Opposition for each record
- Task 2: Sentiment analysis for controversial data
- Task 3 : General overview about entire dataset

TASK 1

- Identify political party as government or opposition
- Method:
  - Using python wikipedia library to get list of parties and their member

Duration	Government
Sep. 24, 1867–Jul. 8, 1872	Conservative Party • Sir John A. Macdonald–1st Ministry
Sep. 3, 1872–Jan. 2, 1874	Conservative Party (1872â1873) • Sir John A. Macdonald–1st Ministry
Sep. 3, 1872–Jan. 2, 1874	Liberal Party (1873â1874)[nb 4] • Alexander Mackenzie–2nd Ministry
Feb. 21, 1874–Aug. 16, 1878	Liberal Party • Alexander Mackenzie–2nd Ministry
Nov. 21, 1878–May 18, 1882	Conservative Party • Sir John A. Macdonald–3rd Ministry
Aug. 7, 1882–Jan. 15, 1887	Conservative Party • Sir John A. Macdonald–3rd Ministry
Apr. 13, 1887–Feb. 3, 1891	Conservative Party • Sir John A. Macdonald–3rd Ministry
Apr. 7, 1891–Apr. 24, 1896	Conservative Party • Sir John A. Macdonald–3rd Ministry (1891) • Sir John Abbott–4th Ministr

TASK 1

- List of the governing and the opposition parties for each parliamentary period

From	To	Government
07/12/1925	02/07/1926	Liberal
07/12/1925	02/07/1926	Conservative
02/11/1926	30/05/1930	Liberal
18/08/1930	14/08/1935	Conservative
09/11/1935	25/01/1940	Liberal
17/04/1940	16/04/1945	Liberal
09/08/1945	30/04/1949	Liberal
29/08/1949	13/06/1953	Liberal
08/10/1953	12/04/1957	Liberal
		Progressive Conservative

TASK 1

- Speechdate: The time for each speech
- Speakerparty: The party for each speech

speechdate	pid	opid	speakeroldr	speakerpos	maintopic	subtopic	subsubtopic	speechtext	speakerparty
2017-12-13	3176068f-62ee-4b2f-aacd-1131201c6db1	214707	The Spea...					It being Wednesday, we will now have the singing of O Canada led by the hon. member for Edmonton Centre.	Liberal
2017-12-13	e915cec9-7164-4846-aeb5-d5bf2e4f3598	214562	Mr. Xavie...		Statemen...	Holiday G...		Mr. Speaker, all the snow we just got is a sure sign that the holidays are just around the corner. On behalf of the Bloc Québécois, I would like to express our	Bloc Québécois
2017-12-13	2f8fed89-2dde-4683-b1f0-03d1f782d96a	229706	Mr. Natha...		Statemen...	Beaches...		Mr. Speaker, the holidays are a time to spend with family and with our community. We have been lucky to celebrate the holidays in Beaches–East York with many	Liberal
2017-12-13	c4b4aef5-eca1-416d-b2b0-ad4a1a2b5fc9	218055	Mr. Kerry ...		Statemen...	Christmas		Mr. Speaker, I have a little Christmas song to sing, even though I am losing my voice:	Conservative
2017-12-13	63894f25-ef7f-45a2-bac2-1a4ceb85b7ad	219363	Ms. Yasm...		Statemen...	Aga Khan		Mr. Speaker, December 13 marks the 81st birthday of His Highness Prince Karim Aga Khan.	Liberal
					Statem...			Mr. Speaker, I would like to join with my wife, Catherine and our family in wishing	

TASK 1

- Solution
  - ➤ Step 1: Look at the "speechdate" of each speech and find out the corresponding duration in list of party for each parliamentary period
  - ➤ Step 2: Check whether the "speakerparty" of this row is the party in corresponding duration in list of party
  - ➤ Step 3: Assign label for each row based on the result on Step 2
    - ✓ Label 1: Government
    - ✓ Label 2: Opposite
    - ✓ Label 3: Unknown

TASK 1

- Step 1.1: Look at "speechdate" of a row

basepk	hid	speechdate	pid	opid	speakeroldr	speakerpos	maintopic	subtopic	subsubtopic	speechtext	speakerparty
4732512	ca.proc.d...	2017-12-13	e63101b2-e7b2-4890-9b4c-be9cf366866f	229559	Mr. Yves ...		Statemen...	Marc-Aur...		Mr. Speaker, here we are in December, the last month of this magnificent year during which we celebrated Canada's 150th anniversary.	Liberal
4732513	ca.proc.d...	2017-12-13	ff7a7f13-0558-4282-8cfc-9bfdfcbf6a1a	216774	Mr. Micha...		Statemen...	Anti-Semit...		Mr. Speaker, last Saturday in Vancouver, pro-Palestinian demonstrators, in a disgusting display of hate, burned Israeli flags, made anti-Semitic slurs, and called	Conservative
4732514	ca.proc.d...	2017-12-13	fa57da4d-d6e1-44ed-9449-b5b1f6361a5c	219935	Ms. Sonia...		Statemen...	Health		Mr. Speaker, we have some of the best and brightest health researchers and health care professionals in Canada. One such visionary is Dr. Arjun Sahgal, a	Liberal
4732515	ca.proc.d...	2017-12-13	897d99d9-377d-43a8-81be-27a22174136e	234515	Mr. Jean ...		Statemen...	The Chris...		Mr. Speaker, the Christmas season is a special time for many people, but a difficult one for some.	Liberal
4732516	ca.proc.d...	2017-12-13	96924c12-efff-4de2-9f9e-578fe5434342	234765	Hon. Erin ...		Statemen...	National ...		Mr. Speaker, 40 years ago, Prime Minister Pierre Trudeau launched a modernization project for the Canadian Armed Forces to acquire a new fighter jet,	Conservative

- Step 1.2: Find out the corresponding duration in list of party

23/05/2011	02/08/2015	Conservative
03/12/2015	11/05/2019	Liberal

- Step 1.3: Compare the "speakerparty" in Step 1.1 with "conservative" in Step 1.2

TASK 1

- Case 1: Government

pk	hid	speechdate	pid	opid	speakeroldr	speakerpos	maintopic	subtopic	subsubtopic	speechtext	speakerparty
940244	ca.proc.d...	1993-04-19	4e3a9e6...	ca.m.4873	Hon. Her...		ORAL Q...	THE ECO...		Mr. Speaker, I have a question for the Prime Minister. His Minister of National Defence has said she can eliminate the deficit in five years. His Minister of the Environment has said	Liberal
940245	ca.proc.d...	1993-04-19	e2215b8...	ca.m.2460	Mr. Edwa...	Chief Gov...	ORAL Q...	THE ECO...		Herbie, you are leading with your chin again.	Progressive Conservative
940246	ca.proc.d...	1993-04-19	1335c5d...	ca.m.8721	Right Ho...	Prime Min...	ORAL Q...	THE ECO...		Mr. Speaker, I do not know if it is entirely accurate that it can be done in five years as the Minister of National Defence says, or in four years as the Minister of the Environment says. I do know that after 25 years of Liberal government it cannot be done by	Progressive Conservative
940247	ca.proc.d...	1993-04-19	4e3a9e6...	ca.m.4873	Hon. Her...		ORAL Q...	THE ECO...		Mr. Speaker, I am surprised that the Prime Minister, in boasting about his record, did not boast about how in only nine years he doubled the public debt to $450 billion, more than happened in all the years since Confederation.	Liberal

24/09/1984	01/10/1988	Progressive Conservative
12/12/1988	08/09/1993	Progressive Conservative
15/11/1993	27/04/1997	Liberal

- Progressive Conservative match to Progressive Conservative

Task 1

- Case 2: Opposite

- New Democratic Party not match to Liberal

TASK 1

- Case 3: Unknown

speechtext	speakerparty	sp
given, they are not accorded any power beyond that which the ordinary citizen of Canada has. House again in Committee on Bill No. 45, Mr. McKENZIE: What struck me as unto incorporate the National Wholesale usual was that a roving committee of that Grocers' Association, and recommending kind should have the right to put people on that the title of said Bill be changed to "An oath and then examine them. But if this Act to incorporate the Canadian Wholesale does not give them any power to swear a Grocers" Association."-(Mr. Sheard)-Mr. witness, I suppose they could send a eom-Boivin in the Chair.mittee around and get statutory evidence On section 12-Committees of Inquiry:without any power at all.		
I understand that when the Bill was last before the Committee, the leader of the Opposition called attention to clause 11 as being undesirable, and suggested it should not have been allowed to pass by the Department of Justice. He also objected to certain features of clause 12. Upon inquiry, I find that clause 11 was struck out before the Private Bills Committee, and clause 12 was	Unionist	

- No information about speackerparty

TASK 1

- Before adding label

basepk	hid	speechdate	pid
4758737	ca.proc.d.2018-1...	2018-10-30	07cf5767-80
4758738	ca.proc.d.2018-1...	2018-10-30	1a1a9dee-c2
4758739	ca.proc.d.2018-1...	2018-10-30	1a1a9dee-c2
4758740	ca.proc.d.2018-1...	2018-10-30	07cf5767-80
4758741	ca.proc.d.2018-1...	2018-10-30	661a3062-86

- After adding label

Role	basepk	hid	speechdate	pid
Government	4758737	ca.proc.d.2018-10-30.10337444	2018-10-30	07cf57(
Government	4758738	ca.proc.d.2018-10-30.10336878	2018-10-30	1a1a9d
Government	4758739	ca.proc.d.2018-10-30.10337508	2018-10-30	1a1a9d
Government	4758740	ca.proc.d.2018-10-30.10336888	2018-10-30	07cf57(
Opposite	4758741	ca.proc.d.2018-10-30.10336896	2018-10-30	661a30

Task 2 : Finding controversial topics

- Method 1 : Using Wikipedia
  - Check if subcat belongs to wikipedia controversial list
  - If controversial topic is mentioned in speechtext atleast twice

Oral Questions	Statement by Members
4927 / 126318 **(3.9%)**	5370 / 43251 **(2.41 %)**

METHOD 2 : NLTK VADER FOR SENTIMENT ANALYSIS

- Sentiment Metric - Positive, Neutral, Negative and Compound.

- How to analyze sentiment based on compound score?
  - 1. Positive sentiment: compound score >= 0.05
  - 2. Neutral sentiment : (compound score > -0.05) and (compound score < 0.05)
  - 3. Negative sentiment : compound score <= -0.05

Oral Questions	Statement by Members
119546 / 126318 **(94.63 %)**	42068 / 43251 **(97.26 %)**

METHOD 3 : COMBINING BOTH APPROACH

- Apply wikipedia method
- Further confirm with Sentiment analysis

Oral questions	Statement by members
4772 / 126318 **(3.77 %)** (wikipedia-155)	5327/ 43251 **(12.31 %)** (wikipedia - 43)

Summary

Method	Oral Questions	Statement by Membes
Wikipedia	4927 / 126318 **(3.9%)**	5370 / 43251 **(2.41 %)**
Sentiment analysis	119546 / 126318 **(94.63 %)**	42068 / 43251 **(97.26 %)**
Combine	4772 / 126318 **(3.77 %)**	5327/ 43251 **(12.31 %)**

ISSUES IN CONTROVERSIAL TOPICS

- Wrong classification by sentiment score
- Controversial data but score is 0.02

"the prime ministers meeting with the president must not turn into an episode of the apprentice onefifth of our workers depend on canadaus trade for their jobs donald trump wants to move those jobs south the liberals are working hard to help him new liberal carbon taxes, payroll taxes, and taxes on small businesses are driving out of our country when will the stop taxing out of canada? "

- Wrong classification by wikipedia
- Not controversial data but score is 0

"that is not true once again, i am sorry that my comments were misinterpreted i know quebeckers love their environment that is why they should be part of our plan, but we do need a national for all quebeckers and "

TASK 3

- Remove duplicated data in entire dataset

	Count	Percentage
Before removing	4749330	100%
After removing	3796847	79.94%

- The final statistic

	Count	Percentage
Government	1434816	37.79%
Opposite	1890925	49.80%
Unknown	471106	12.41%
Total	3796847	100%

TASK 3

Compare *labelled data* **vs** **unlabelled data**

Pie chart: 12,41% (unlabelled data), 87,59% (labelled data)

- labelled data
- unlabelled data
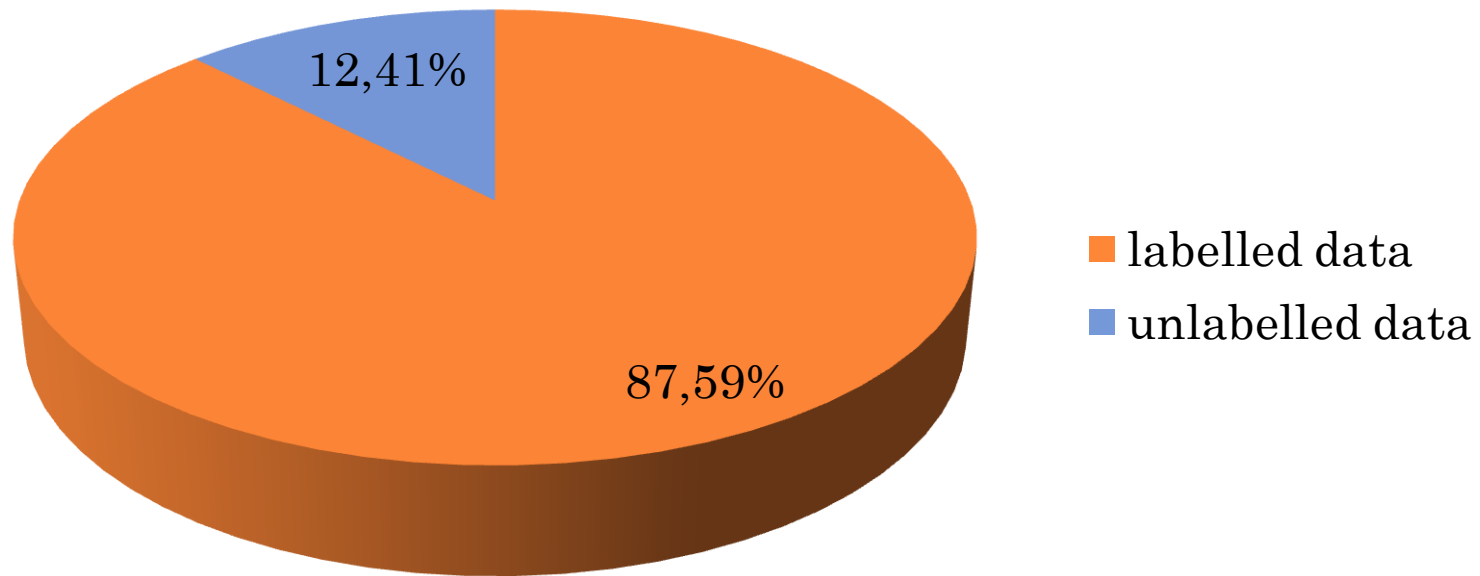
TASK 3

General information for Oral Questions
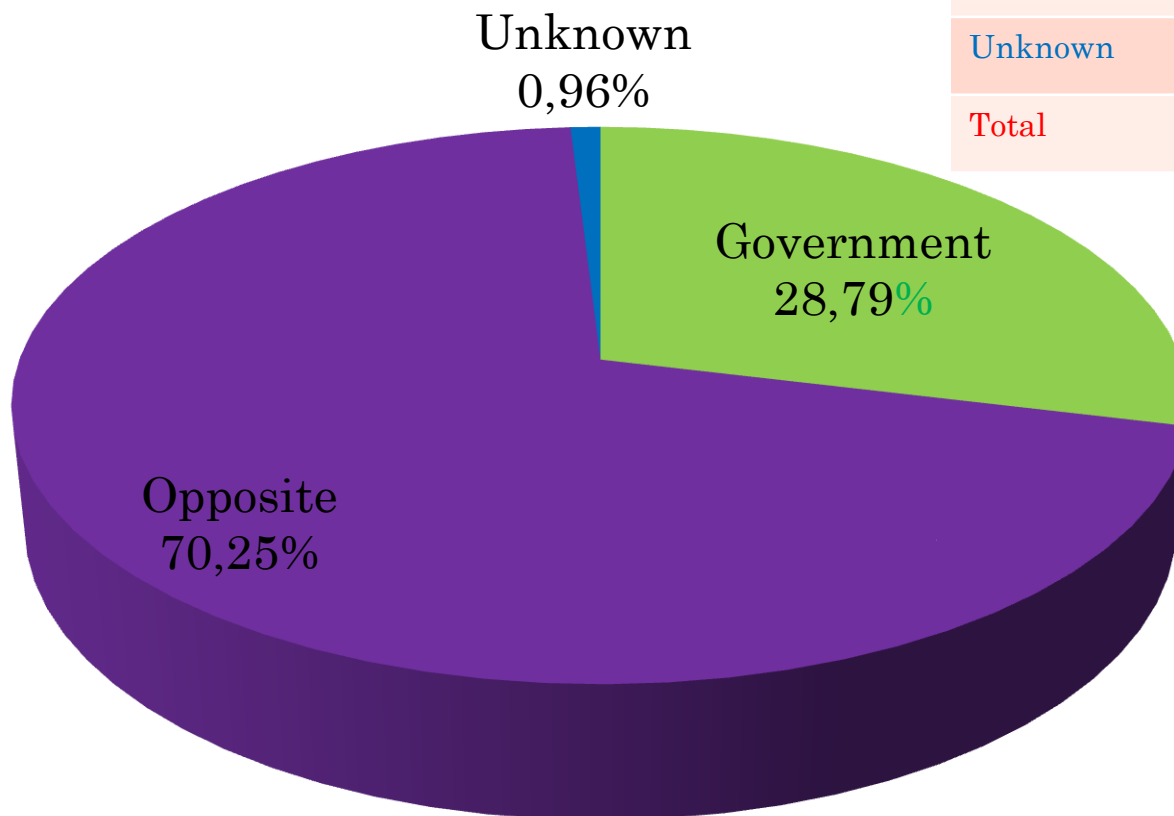
	Count	Percentage
Government	36367	28,79%
Opposite	88733	70,25%
Unknown	1218	0,96%
Total	126318	100%

Unknown
0,96%

Government
28,79%

Opposite
70,25%

General information for Statement By Member

	Count	Percentage
Government	17624	40,75%
Opposite	24723	57,16%
Unknown	904	2.09%
Total	43251	100%

Unknown
2,09%

Government
40,75%

Opposite
57,16%
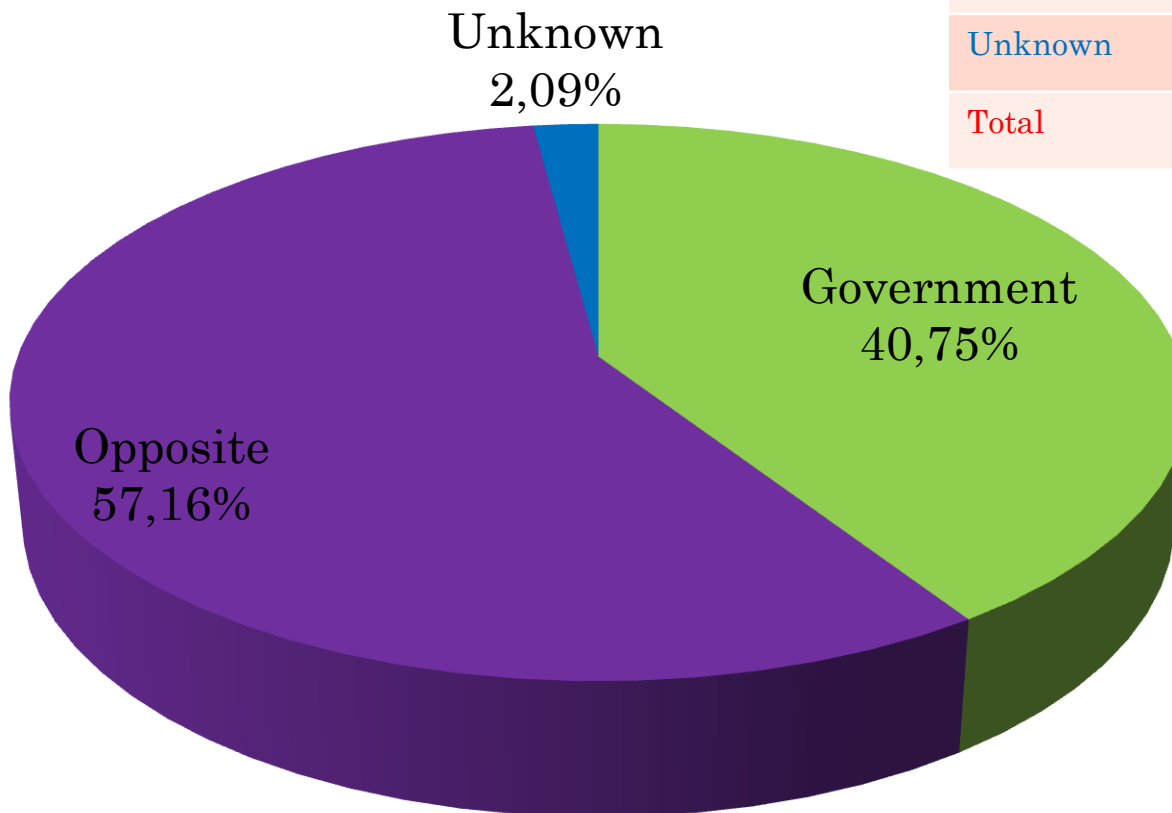
TASK 3

- Top 10 parties based on these speakerparty

	Count	Percentage
Liberal	1.347.861	35,50%
Progressive Conservative	655.773	17,27%
Conservative (1867-1942)	352.607	9,29%
New Democratic	274.843	7,24%
Conservative	167.829	4,42%
Co-operative commonwealth federation (c.c.f.)	103.316	2,72%
Bloc Québécois	65.447	1,72%
Social Credit	48.819	1,29%
National Government	41.811	1,10%
Reform	38.092	1,00%
others	700.449	18,45%
total	3.796.847	100%

TASK 3

- Top 10 main topic in entire dataset

	Count	Percentage
government orders	605206	15,94%
oral question period	397990	10,48%
routine proceedings	182673	4,81%
questions	166741	4,39%
supply	128928	3,40%
oral questions	126318	**3,33%**
ways and means	67823	1,79%
statements by members	43251	**1,14%**
external affairs	40814	1,07%
private members' business	37850	1,00%
other	1.999.253	52,66%
total	3.796.847	100%

QUESTION

What we should do with

## unlabelled data ?

SOLUTION

- Collection information from other corresponding cell (main topic, sub topic, speakername) to define the label

- Issue

| speakeroldr	speakerpos	maintopic	subtopic	subsubtopic	speechtext	speakerparty	speakerridir	speakername
nan	nan	SUPPLY-AUDIT...	nan	nan	My hon. friend (Mr. Wallace) will find no cause of complaint on that score.	nan	nan	The MINISTER OF FINANCE.
nan	subtopic	WAYS AND ME...	nan	nan	["The hon. gentleman is just as well aware as I am that Mr. FIELDING.\nthe report of the Auditor General is not at all under the control of the government. The government is not at all answerable for the delay of the Auditor General's Report. With regard to the report of the Minister of Customs, the House I am sure will learn with regret that the Minister of Customs has been called away from Ottawa by sickness in his family. But he will be here to-morrow, I hope.\nOn motion of the Prime Minister, the House adjourned at 5.15 p.m.", 'Tuesday, February 12, 1901.']	nan	nan	nan
nan	nan	WAYS AND ME...	nan	nan	That this House will, to-morrow, resolve itself into a committee to consider of the Ways and Means for	nan	nan	The MINISTER OF FINANCE (Ho
nan	stagedire...	nan	nan	nan	The MINISTER OF FINANCE (Hon. Mr. Fielding) presented a Message from His Excellency the	nan	nan	nan

Recall to overview of entire dataset

Compare labelled data vs unlabelled data

- labelled data
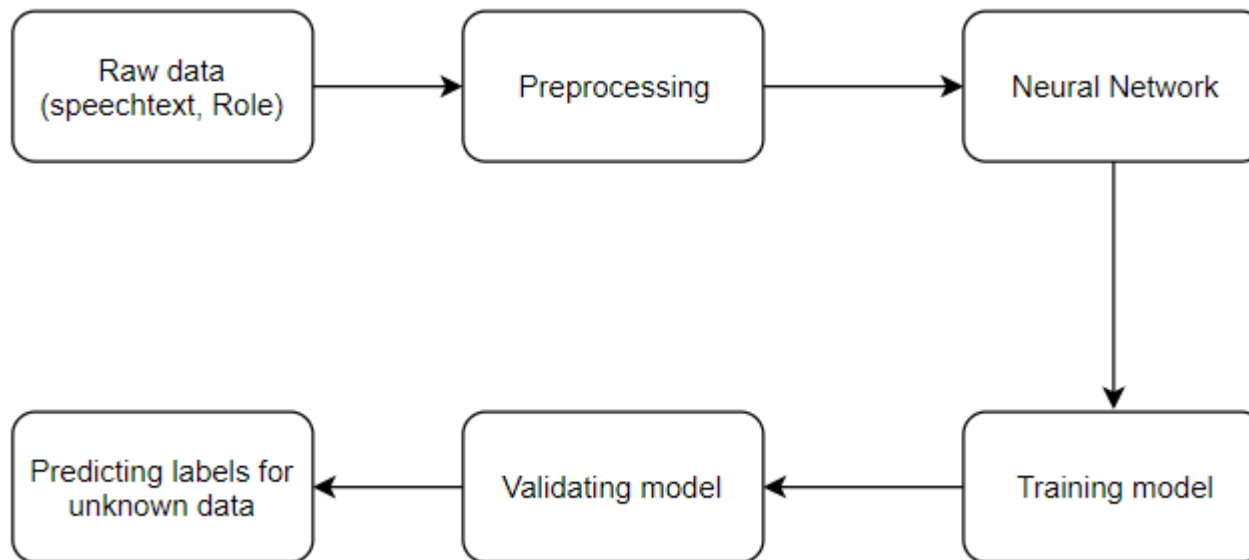- unlabelled data

12,41%

87,59%
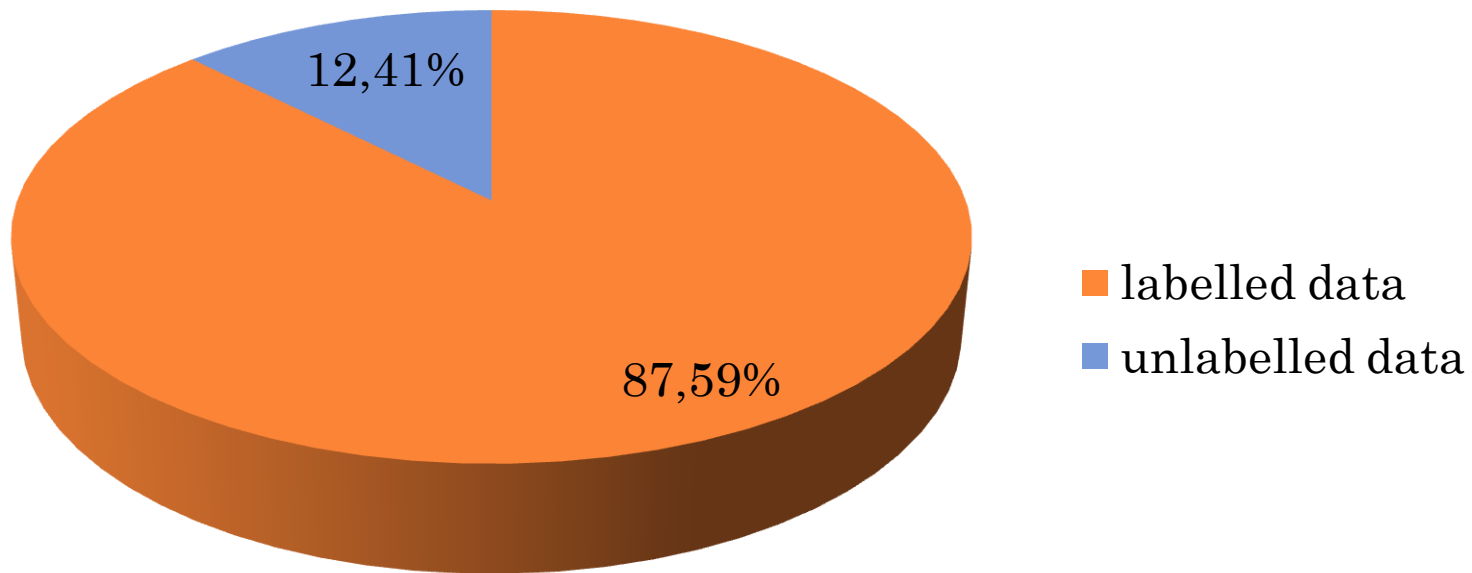
DEEP LEARNING FOR TEXT/DOCUMENT CLASSIFICATION

- Overview for Text/Document/Label classification
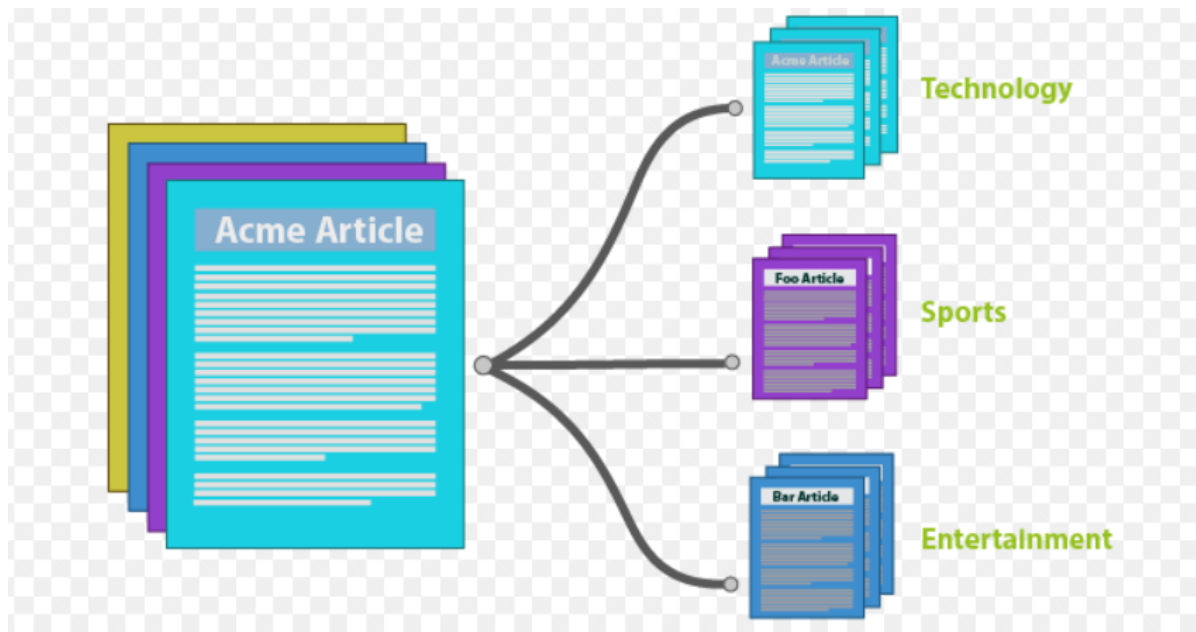
DEEP LEARNING FOR TEXT/DOCUMENT CLASSIFICATION

- General Process

RAW DATA

- Training & validation set

Role	basepk	hid	speechdate	pid
Opposite	01/03/1900	ca.proc.d.1901-0...	11/02/1901	57ed8b95-0f38-4...
Government	06/03/1900	ca.proc.d.1901-0...	11/02/1901	83c162c1-32ed-...
Government	10/03/1900	ca.proc.d.1901-0...	11/02/1901	a8b2b613-dd5c-...
Government	12/03/1900	ca.proc.d.1901-0...	11/02/1901	a8b2b613-dd5c-...
Government	17/03/1900	ca.proc.d.1901-0...	11/02/1901	a8b2b613-dd5c-...
Government	18/03/1900	ca.proc.d.1901-0...	11/02/1901	a8b2b613-dd5c-...

- Testing set

| Role	basepk	hid	speechdate	
Unknown	55	ca.proc.d.1901-02-11.1	11/02/1901	n
Unknown	56	ca.proc.d.1901-02-11.1.1	11/02/1901	in
Unknown	57	ca.proc.d.1901-02-11.1.2.1	11/02/1901	n
Unknown	59	ca.proc.d.1901-02-11.2.1	11/02/1901	u
Unknown	60	ca.proc.d.1901-02-11.2.2	11/02/1901	in
Unknown	62	ca.proc.d.1901-02-11.2.4.1	11/02/1901	n

PREPROCESSING

- Remove cliches
- Remove stopwords
- Remove punctuation
- Tokenization
- Vectorization
- Embedding

NEURAL NETWORK
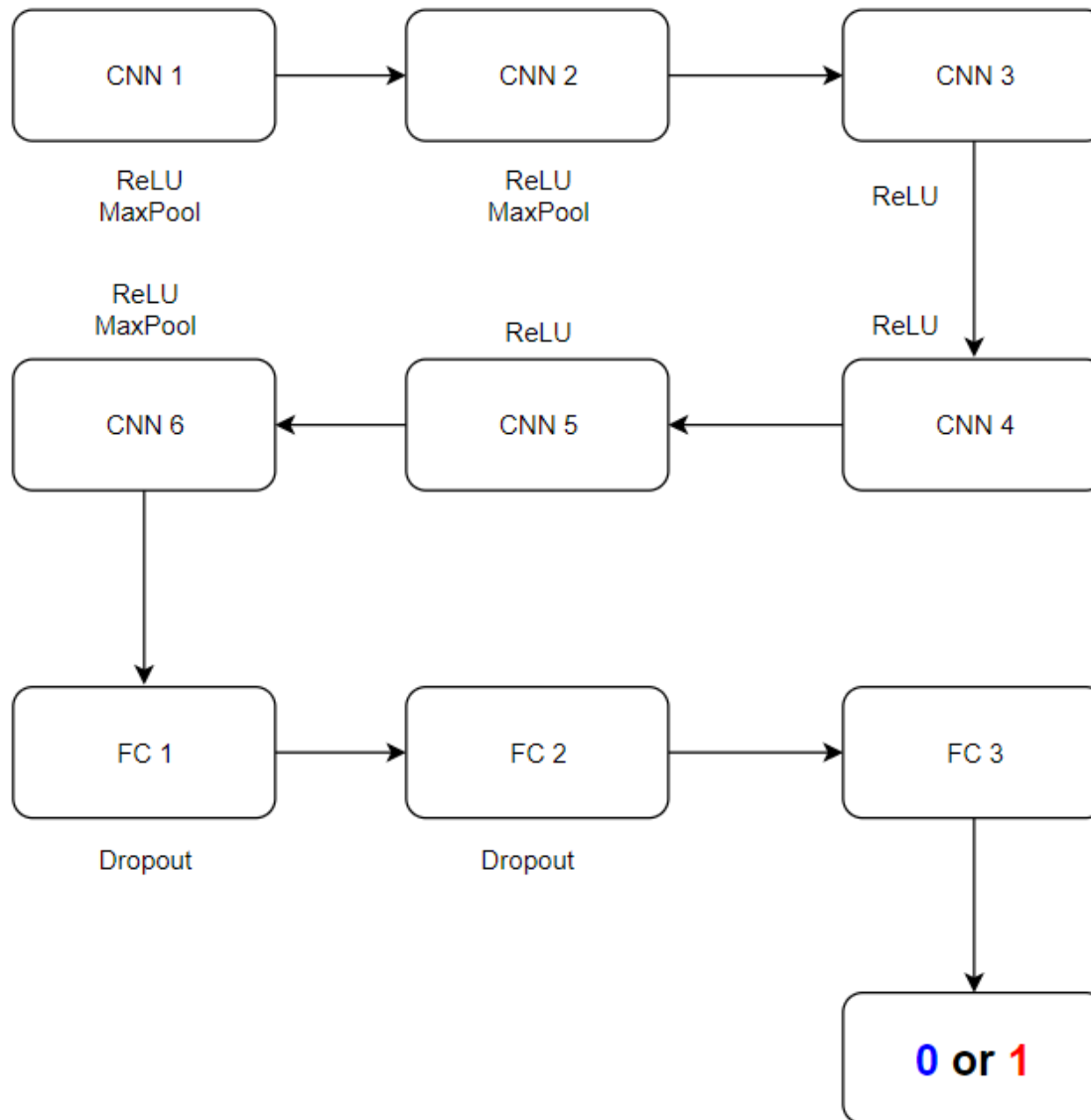
CONFIGURATION

- Training set: 18000 rows (9000 for Government, 9000 for Opposite)

- Validation set: 2000 rows (1000 for Government, 1000 for Opposite)

- Test set: all of unknown rows (471106 rows)

- Batch size: 32

- Epoch: 1

- Learning rate: 0.001

- Optimization function: Adam

- Loss function: Cross Entropy

VISUALIZATION

- Accuracy

Epoch 1/1. Loss: 0.6972806453704834, Accuracy: 0.5211693644523621

- Prediction

[1 1 1 1 1 1 1 1 1 1 0 1 1 0 0 0 0 1 1 1 1 1 1 1 0 1 0 1 0 0 1 0]
[1 1 1 0 0 1 1 1 0 1 0 1 1 1 0 1 1 1 0 0 1 1 1 1 0 1 1 0 1 1 0 0]
[1 0 0 0 0 1 0 0 1 0 0 1 1 0 0 1 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1]   **1 batch size**
[1 1 1 1 0 0 1 1 1 0 1 0 1 1 0 1 1 0 1 1 1 1 1 1 1 1 0 1 1 0 1 1]
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 1 1 0 0 1 1 1 1 1 1 1 1 0 0]

- 0: Government
- 1: Opposite