Automated Multi-Label Classification

Code Gladiators 2020

Team Name : AI Enthusiasts

Team Member: Anand P V
M. Tech CSE, IIT Bombay
VP of Technology, IIA Pvt Ltd.

Executive Summary

• Innovative way to do news classification with a **unique combination of LDA, NMF and One Vs Rest classifier**.

•LDA is a **probabilistic generative process** adapted to different kind of analysis.

• NMF is a **linear-algebraic meth**od to calculate the value of keywords in the document and assign document to topic which keyword has maximum value.

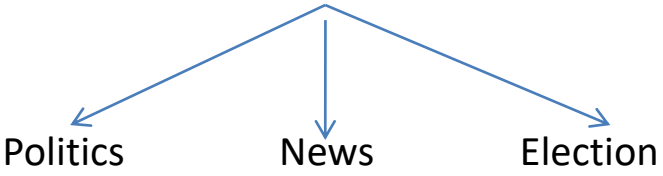
• Google's **300-dimensional word vectors** for a vocabulary of 3 million words, which are trained on around 100 billion words from a Google News dataset is used to compute word distance metrics.

• Category Tree information **embedded in HTTP link** is extracted using interesting properties of word vectors.

• **Significant words in "Description"** is correlated with categories in given enriched Category Tree by TIL.

• **Breadth-First Search** of category tree is done to find out the best category tree allocation based on W2V

• **One Vs Rest classifier** is used to resolve classification ambiguities.



• The **outcome of NMF and LDA are normalized and conjugated** before figuring out the topic.

Problem at hand

TECHGIG CODE GLADIATORS 2020

- The problem at hand is Automated Multi-Label Classification .

- The requirement is to **classify the news category based on the collated news article data containing text and URL.**

- News can belong to one or more category based on content of the news.

- For example: *News about sports personality may belong to the Sports category and Entertainment category too.*

- There are certain entities/keywords in the news article which can help to determine category of the news.

- All the URL's visited on TIL network. Given set of URL's, their short/Long description, participants have to classify them across multiple categories.

Tasks:

- **Identification of Keywords :** The content which can help in categorizing the news to different categories

- **Formation of category tree** on the basis of the identified keywords with ML.

Category Tree Depiction

Solution

• As the dataset is very huge, the task to label articles in to a category hierarchy would take huge amount of time. The aim is to reduce the manual work to the extend possible.

• When I used the state-of-the-art Topic-Modelling technique **LDA (Latent Dirichlet Allocation), it gave high number of wrong classification. Hence NMF (Non Negative Matrix Factorization) is also used along with Multi Class Binarizer to refine results.**

- **As LDA, NMF and MC-Binarizer resulted in fairly high number of miss-classification, we have found a unique way to numerically combine LDA matrix output with factorized output of NMF, post-normalization and then match with MC-Binarizer to improve accuracy.**

• **LDA is a probabilistic model and NMF a linear algebraic model** used to find common keywords that appear in number of documents and assign those to the topics.

• For example: consider you have around 100000 doc's and specify n_topics = 10 while applying LDA function. LDA will sort all doc's into 10 and assign keyword's to the topics.

Topic 1 : 0 – 9999
Topic 2 : 10000 – 19999
-
-
Topic 10 : 90000 – 99999

• After finding keywords I manually assigned keywords to the Topics. Then, assigned keywords to the topic which has maximum value.

• For example: if we have keywords "bjp, election, modi", then such document will be assigned to topic politics and election.

• After that used **One Vs Rest classification algorithm with Logistic Regression** to resolve ambiguities, for instance
- the bjp, modi keywords into Politics
- then election keyword into Election.

Solution

- Though **LDA and NMF method**s provide the necessary categorization, it **wont be able to provide the multi-level category-tree differentiation** of data, as required by TIL enriched category tree.

- Hence, I have **extracted the category-tree information hidden in the sequence of URL text**, corresponding to each article.

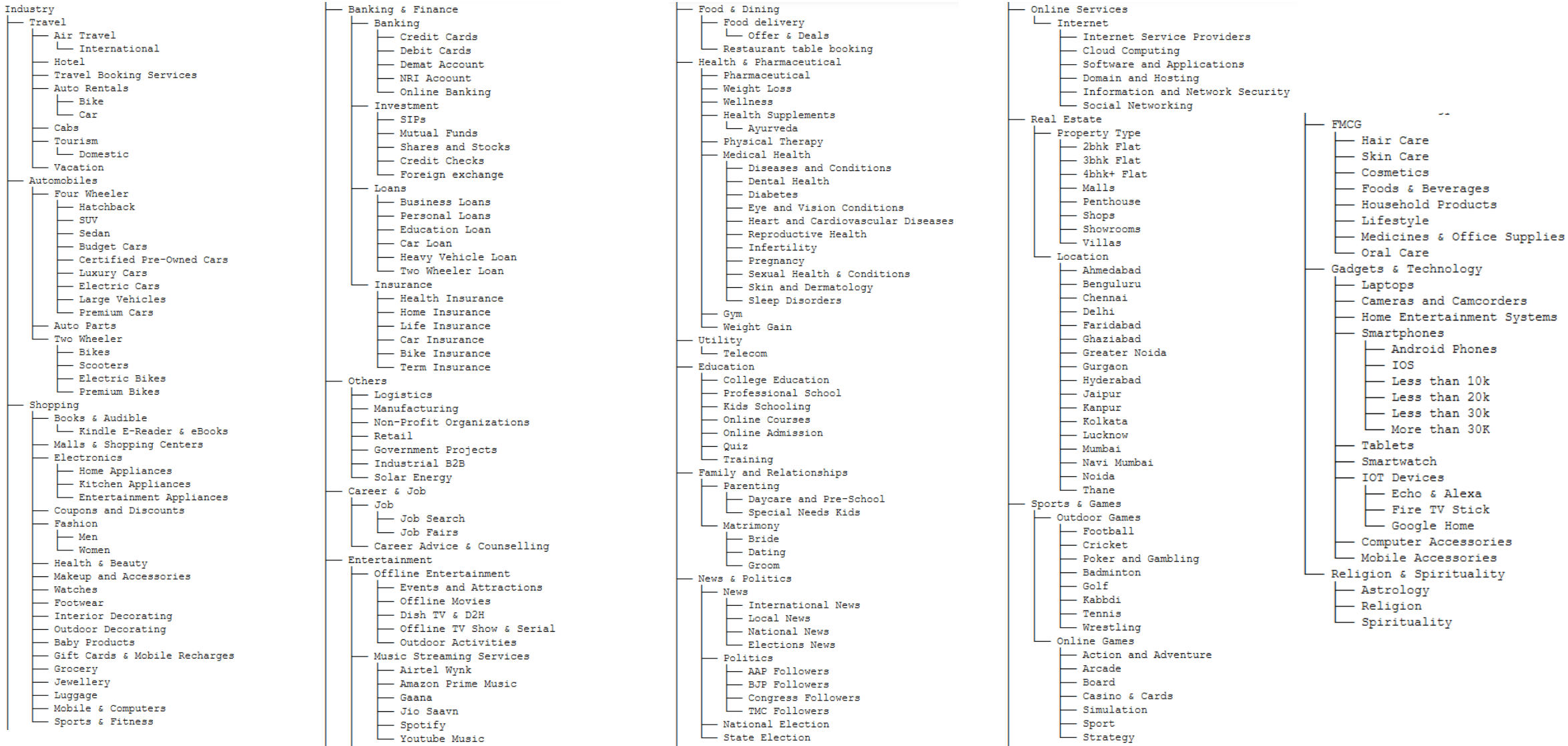
- A custom algorithm which **peruses GoogleNews vectors and category tree to parse each URL** is coded. The corresponding confidence score of category prediction is computed using a mathematical formula based on the numerical distribution of values in the list. Saved the (id, tree, **confidence**) information into CSV file.

- Another **deep algorithm is used to analyze each article text correlated to each category node**. Multi-word categories are intelligently handled not to hike up the similarity scores. This function is the **core** of the solution. Here also, I have saved the (id, tree, confidence) information into CSV file.

- Next step, I have **merged the Article Description Classifier with URL Classifier as first fallback and LDA-NMF Classifier as second fallback. The confidence score is used to decide the relative relevance** of the classification method. Article Description Classifier is considered as the final fallback as it is the most analytical.

Solution Details

AnyTree package in Python is used to create, manipulate and traverse the category tree information given as input. The below tree structure is programmatically generated using the given csv input and visualized using **graphviz**.



Solution Demo

TECHGIG CODE GLADIATORS 2020

LDA Output:

These are the Common Keywords extracted from docs using LDA and labelled topics,

topic_theme	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10	Word 11	Word 12	Word 13	Word 14	Word 15	Word 16	Word 17	Word 18	Word 19
Finance/OTT APPS/News/Business/Location	announce	budget	finance	end	join	bengaluru	follow	leave	lakh	game	telegram	centre	business	fund	project	important	likely	maker	channel	question
Sports/Medical Health/Politics	india	new	video	test	series	like	zealand	south	indian	global	claim	cricket	right	raise	australia	odi	future	virus	brand	run
Politics/Entertainment	mumbai	medium	social	include	maharashtra	fall	couple	ministry	deepika	padukone	son	priyanka	real	ban	red	nation	pro	unveil	motor	town
News/Entertainment	high	news	life	movie	direct	director	award	service	international	make	close	past	washington	base	fight	year	send	level	allow	finally
Song/News	report	woman	song	plan	music	assembly	face	pakistan	accord	night	late	head	away	person	kerala	shoot	girl	check	charge	rape
Kids Schooling/Entertainment	make	hold	child	use	story	popular	talk	special	office	song	kid	recent	baby	present	statement	ram	box	nursery	watch	sri
Banking & Investment/News/Education	post	city	appear	just	university	league	force	want	air	investment	set	action	long	break	day	newscrab	live	know	amid	premier
Entertainment	film	actress	actor	khan	bollywood	recently	look	release	fan	love	star	time	upcoming	share	kapoor	year	picture	set	allegedly	great
Sports & Entertainment/Death	come	day	make	capital	continue	lead	sharma	expect	player	death	final	feature	remain	say	kohli	reach	virat	add	like	soon
News & Politics	time	february	trump	president	pradesh	tnn	donald	hit	visit	uttar	feb	increase	write	hand	book	spot	army	super	beat	try
Health & Pharameutical/OTT APPS	day	coronavirus	china	people	group	kill	ali	road	family	industry	near	outbreak	rule	celebrate	early	hour	instagram	injure	travel	index
Banking/Investment/Economy	sunday	cent	bank	crore	market	new	share	trade	delhi	stock	price	session	point	gold	say	friday	development	quarter	peer	firm
Government Projects/Improvement/Finance	new	launch	thursday	year	government	bring	delhi	united	rise	growth	help	sector	price	flipboard	financial	airport	oil	station	record	corporation
OTT APPS/Smartphone	ist	update	feb	search	company	jan	india	pti	new	app	million	pay	smartphone	chinese	sale	december	policy	low	datum	user
Education/Entertainment/Economy	student	big	match	photo	image	turn	source	wwe	twitter	world	event	economic	performance	mark	sign	thing	different	college	year	spread
Politics/News/Election	delhi	new	court	say	india	monday	leader	month	national	tuesday	wednesday	feb	ani	issue	ask	order	seek	state	supreme	change
Politics	minister	say	chief	delhi	government	party	friday	bjp	state	today	congress	union	act	home	protest	modi	prime	thursday	citizenship	narendra
Job/Health	work	say	flipboard	meet	board	health	india	department	power	official	need	indian	offer	hospital	tax	target	kumar	die	january	candidate
News & Sports/Entertainment	team	world	win	indian	boss	open	house	bigg	good	season	second	way	cup	play	india	support	champion	old	round	drop
News/Education	police	saturday	man	case	singh	week	arrest	district	school	file	chennai	kolkata	role	officer	tamil	kejriwal	say	arvind	aap	punjab

NMF Output:

These are the Common Keywords extracted from docs using NMF and labelled topics,

topic_theme	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10	Word 11	Word 12	Word 13	Word 14	Word 15	Word 16	Word 17	Word 18	Word 19
Politics/Election	delhi	party	election	assembly	bjp	congress	leader	kejriwal	arvind	poll	aam	feb	aap	aadmi	janata	violence	candidate	bharatiya	seat	capital
Entertainment	film	actor	khan	release	movie	ali	star	bollywood	role	trailer	director	kapoor	sara	box	office	love	aaj	audience	maker	starrer
OTT APPS/ News/ Banking & Investment	telegram	join	app	importantâ	channelâ	investment	download	business	update	finance	httpstmeinvestmentguruindiaâ	th	channel	ringtone	news	sitharaman	report	budget	company	view
Music	song	music	fan	checkout	bhojpuri	sing	lyric	video	singer	sung	punjabi	movie	singh	haryanvi	star	ke	hit	feature	gana	holi
Banking & Investment	cent	crore	company	share	rs	bank	market	price	profit	quarter	stock	session	trade	report	rise	index	growth	end	lakh	peer
News	police	man	woman	district	city	station	area	person	girl	incident	night	village	accuse	people	arrest	officer	road	crime	case	force
News & Sports	india	world	team	cricket	zealand	series	match	cup	test	virat	game	kohli	odi	australia	captain	player	photo	league	win	source
Politics/ Budget	minister	union	modi	narendra	budget	finance	home	sitharaman	leader	shah	amit	singh	saturday	thackeray	kejriwal	meeting	cabinet	nirmala	chief	thursday
Offline Entertainment	actress	bollywood	post	medium	picture	share	fan	today	video	look	kapoor	photo	tv	instagram	newscrab	padukone	fashion	actor	khan	beauty
Improvement	year	growth	couple	month	world	decade	number	sale	award	industry	age	wwe	december	start	work	end	accord	plan	celebration	record
News & Politics	act	citizenship	amendment	caa	protest	register	nrc	law	resolution	bagh	congress	people	opposition	citizen	support	leader	country	bjp	protester	population
Health & Pharmateutical	coronavirus	people	china	outbreak	health	death	case	country	virus	toll	report	number	city	world	spread	authority	hospital	infection	wuhan	rise
Relationship/News	day	valentines	love	today	week	test	news	couple	temperature	group	office	valentine	celebrate	celebration	trophy	occasion	box	army	event	transfer
Education/Goverment Projects	government	state	school	department	plan	centre	budget	tax	scheme	policy	education	issue	decision	congress	kerala	hospital	district	service	pradesh	secretary
News	jan	ist	search	tnn	image	mumbai	pti	pm	feb	news	photo	ani	bengaluru	timesofindiacom	chennai	file	man	karnataka	hyderabad	hold
Politics	president	trump	visit	donald	feb	washington	melania	modi	states	deal	trumps	ahmedabad	india	trade	narendra	lady	pm	congress	house	gandhi
Kid Education/Education	story	child	kid	baby	song	rhyme	nursery	poem	january	watch	children	telugu	malayalam	html	headheadhtml	school	girl	parent	woman	animate
News	court	case	convict	order	plea	petition	death	murder	rape	justice	file	issue	judge	gangrape	bench	sentence	bail	row	jail	thursday
Primary Education/Secondary Education	student	university	jnu	school	jawaharlal	violence	college	campus	protest	class	nehru	attack	teacher	education	board	examination	jamia	exam	institute	padukone
Entertainment	time	boss	bigg	house	contestant	season	week	episode	shukla	news	sidharth	khan	family	riaz	reality	gill	asim	actor	lot	fan

Combined LDA and NMF output:

After finding top keywords from both LDA & NMF. Then found probability of documents in LDA and NMF. Then I combined both models .

topic_theme	Politics/Election	Entertainment	OTT APPS/ News/ Banking & Investment	Music	Banking & Investment	News	News & Sports	Politics/ Budget	Offline Entertainment	Improvement	...	Health & Pharameutical/OTT APPS	Banking/Investment/Economy	Government Projects/Improvement/Finance	OTT APPS/Smartphone	Education/Entertainment/Economy	F
0	0.000000	0.000000	0.000000	0.000000	0.000000	0.048856	0.000000	0.431489	0.000000	0.000000	...	0.002646	0.002564	0.002643	0.002351	0.002524	
1	0.000000	0.000000	0.000000	0.131736	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.002054	0.001990	0.002038	0.001777	0.001945	
2	0.000000	0.000000	0.196065	0.000000	0.000000	0.013570	0.030205	0.000000	0.000000	0.000000	...	0.003178	0.003079	0.003187	0.002868	0.003045	
3	0.000000	0.000000	0.346378	0.000000	0.026441	0.000000	0.036412	0.000000	0.014843	0.014171	...	0.002054	0.001990	0.002038	0.049285	0.045944	
4	0.004812	0.164271	0.000000	0.000000	0.000000	0.170582	0.018000	0.089539	0.206756	0.000000	...	0.002646	0.002564	0.002643	0.002351	0.002524	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
410402	0.032685	0.000000	0.000000	0.000000	0.015940	0.809545	0.000000	0.000000	0.001605	0.000000	...	0.002504	0.002426	0.112853	0.140454	0.002385	
410403	0.056797	0.000000	0.000000	0.000000	0.073221	0.056240	0.077162	0.029158	0.041263	0.000000	...	0.003660	0.003546	0.003680	0.003336	0.003516	
410404	0.015214	0.000000	0.000000	0.000000	0.069951	0.000000	0.560531	0.017679	0.000000	0.000000	...	0.003403	0.003297	0.003417	0.003087	0.214462	
410405	0.028273	0.000000	0.000000	0.000000	0.073449	0.000000	0.003680	0.008588	0.000000	0.000000	...	0.184143	0.004168	0.248890	0.091277	0.004145	
410406	0.000000	0.008776	0.000000	0.000000	0.608636	0.085395	0.000000	0.009508	0.000000	0.005884	...	0.147276	0.142674	0.076988	0.003087	0.073664	

410407 rows × 40 columns

LDA-NMF Result:

After combining the models I assigned topic to the document which topic has maximum percentage.

	question_text	id	dominant_topic	dominant_topic_theme
0	bhopal the widespread protests against caa are proof that indian voters especially the youth will not be taken for a ride any longer believes madhya pradesh chief minister kamal nath in an exclu	5177	21	Sports/Medical Health/Politics
1	heres presenting popular children nursery story krishna and kaliya sri krishna for popular children stories kids songs children songs children poems baby songs baby rhymes kids nursery rhy	2882	25	Kids Schooling/Entertainment
2	norwich city v tottenham hotspur premier league hello and welcome to the epl transfer news roundup for the day here are the top stories of the day surrounding the erling braut haaland joins dortm	2675	26	Banking & Investment/News/Education
3	business news › news › politics and nation ›winter chill grips north india winter chill grips north india et online and agencies dec am ist dal lake begins to freeze kashmir woke u	2676	35	Politics/News/Election
4	the bollywood fraternity on friday expressed pain and shock on the news of tv actor kushal punjabis untimely death kushal committed suicide at his home in mumbais bandra area late on thursday night	3875	36	Politics
...	...	...	...	...
410402	aligarh uttar pradesh feb ani suspension of mobile internet services on thursday was extended till february in aligarh district following recent clashes between police and anticit	163675	5	News
410403	recalling his long association with rss ideologue p parameswaran who passed away last week rss sarsangh chalak mohan bhagwat here on wednesday said parameswaran was a model swayam sewak who very mu	166121	36	Politics
410404	jimmy butlers miami heat side have struggled of late match details fixture vs date time friday february pm et venue americanairlines arena miami fl last game results dallas	165505	6	News & Sports
410405	rome feb sputnikani the death toll from coronavirus infection in italy has risen to with the addition of three more victims in the northwest of the country on thursday angelo borre	162919	11	Health & Pharmateutical
410406	by keeping traffic management its top priority ahmedabad urban development authority auda aims to finish construction of four flyovers by with a project cost of rs cr the flyovers on sp r	168359	4	Banking & Investment

Data Pre-processing

• In the dataset we have large number of dataset so we cleaned that before doing topic modelling.

• First problem I encountered is we have some missing content in the long description so I replaced the content with values in the short description.

• And made the description column NAN Free.

• Then the main problem in the dataset is languages since LDA only work with English Language.

• So I used **Google Translate API** to translate all languages into English.

• After that removed stopword's, punctuation, brackets and numbers.

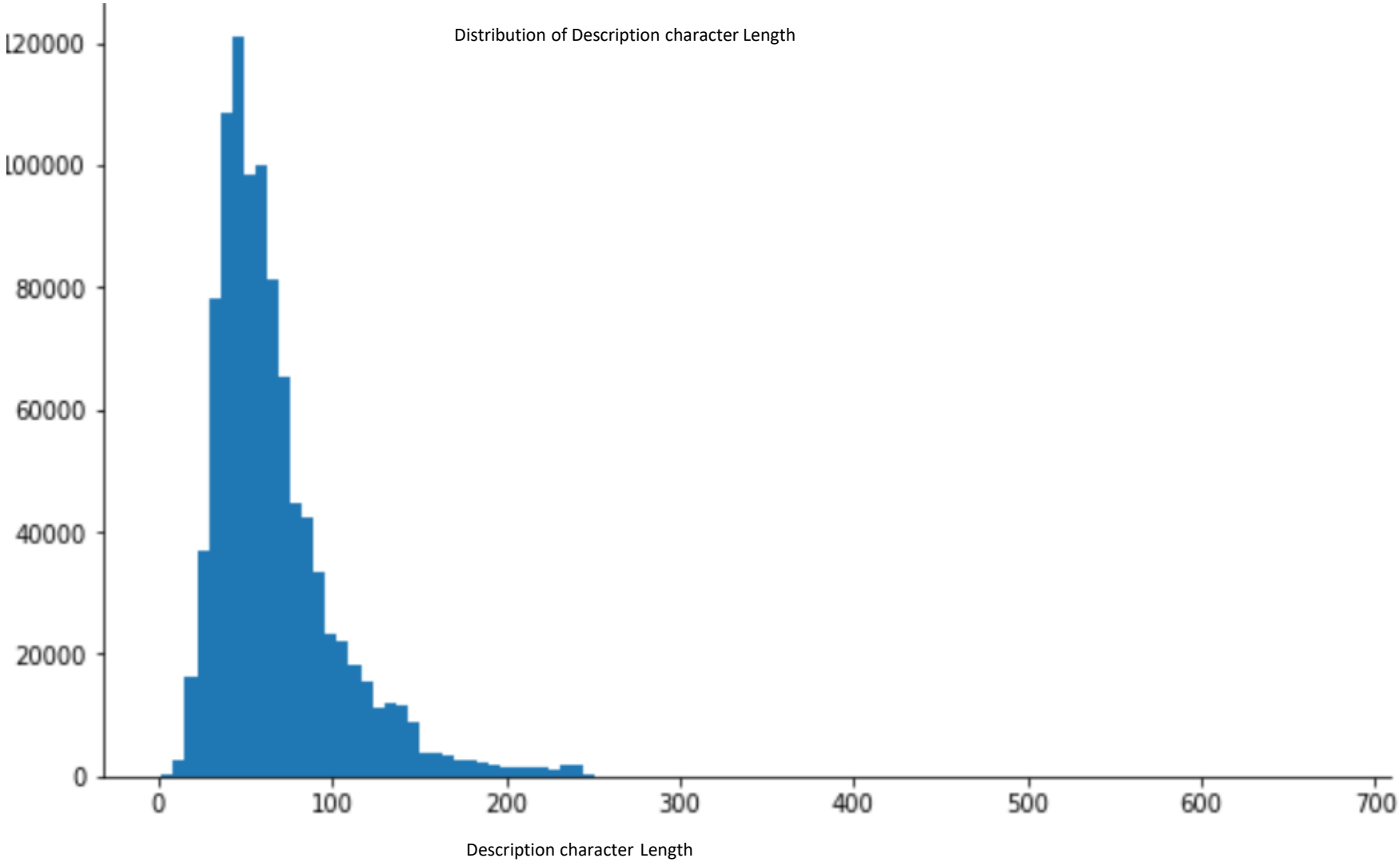
• Lowercase and Lemmatization using Spacy then removing "PRON" & Stemming

• For word vector similarity computation, lemmatization and stemming are avoided as the Google Word vectors are not trained on lemmatized English words.

• The category tree node words are tweaked because multi-words categories having different meanings are combined into similar meaning words so that distance metrics won't go awry. This can easily be mapped back to the original category info. The meaning of the tweaked node names are also meaningful, nevertheless.

Exploratory Data Analysis:

Using Exploratory Data Analysis, visualized how the descriptions are in general,



Topic Modelling with LDA and NMF

- **Latent Dirichlet Allocation (LDA)** and **Non Negative Matrix Factorization** takes only input as a document-term matrix

- So I calculated document matrix for every documents. Here is my example document matrix,

	football	player	stay	safe	most	hate	nfl	team	greatest	political	leader
question 1	1	1	1	1	0	0	0	0	0	0	0
question 2	1	0	0	0	1	1	1	1	0	0	0
question 3	0	0	0	0	0	0	0	0	1	1	1

each row is a document ,each column is word , we label “0” if document doesn’t contain word otherwise label as “1”.

LDA & NMF Output

Top words extracted from all the documents using LDA & NMF.

	Topic # 01	Topic # 02	Topic # 03	Topic # 04	Topic # 05	Topic # 06	Topic # 07	Topic # 08	Topic # 09	Topic # 10
0	queensland	rural	sydney	canberra	wa	police	australian	government	south	trump
1	court	two	world	north	calls	death	australia	says	election	adelaide
2	woman	90	china	perth	new	say	day	found	news	seconds
3	nsw	attack	win	coast	call	missing	country	donald	tasmania	years
4	indigenous	crash	cup	2015	labor	children	one	accused	says	national
5	charged	car	business	afl	funding	nrl	hour	us	water	family
6	child	police	market	gold	nt	darwin	tasmanian	people	victoria	drug
7	power	women	australia	record	council	john	test	trial	top	near
8	farmers	dead	final	four	hobart	west	year	new	take	hospital
9	murder	health	home	interview	qld	cattle	budget	guilty	life	park
10	live	christmas	league	show	royal	road	says	png	residents	energy
11	police	killed	melbourne	cyclone	support	search	change	ahead	northern	new
12	sex	nsw	hit	2016	bill	coal	new	markets		court
13	school	house	turnbull	year	aboriginal	river	violence	week	hill	million
14	federal	second	first	president	tax	2017	changes	super	east	action
15	weather	driver	city	wa	review	new	climate	shows	train	jail
16	victorian	five	young	michael	rise	youth	first	study	security	newcastle
17	port	service	rugby	sport	media	community	laws	push	new	mark
18	island	shooting	wins	regional	deal	body	make	podcast	grand	chief
19	abuse	students	open	risk	housing	fire	time	art	inquest	appeal

LATENT DIRICHLET ALLOCATION (LDA)

	Topic # 01	Topic # 02	Topic # 03	Topic # 04	Topic # 05	Topic # 06	Topic # 07	Topic # 08	Topic # 09	Topic # 10
0	interview	seconds	police	new	fire	abc	rural	charged	council	court
1	michael	90	missing	zealand	house	weather	news	murder	says	accused
2	extended	business	probe	laws	crews	sport	nsw	crash	water	murder
3	david	sport	search	document	threat	news	national	woman	govt	faces
4	james	weather	investigate	hospital	destroys	entertainment	qld	death	us	front
5	john	news	hunt	year	school	business	podcast	car	plan	told
6	nrl	closer	death	home	home	market	reporter	stabbing	australia	charges
7	ivan	confidence	car	deal	blaze	analysis	country	two	report	case
8	matt	exchange	shooting	centre	suspicious	talks	nrn	assault	back	hears
9	nathan	analysis	officer	york	warning	speaks	hour	sydney	closer	drug
10	chris	friday	crash	president	factory	wild	health	fatal	health	high
11	luke	small	seek	gets	season	breakfast	sach	trial	call	sex
12	andrew	wild	arrest	opens	sydney	learning	drought	killed	urged	appeal
13	smith	chamber	fatal	chief	damages	friday	tasmania	attack	hospital	alleged
14	tim	bad	find	named	residents	report	ntch	jailed	wa	assault
15	scott	good	assault	ceo	killed	891	doctors	teen	wins	fronts
16	peter	market	drug	get	control	darwin	quarter	found	australian	trial
17	mark	pm	found	life	destroyed	wednesday	tas	guilty	funding	child
18	matthew	thursday	body	mayor	ban	radio	friday	driver	calls	stabbing
19	shane	monday	station	rules	woman	thursday	monday	shooting	budget	charge

NON NEGATIVE MATRIX FACTORIZATION (NMF)

Evaluate LDA model Performance with Perplexity and Log-likelihood

- **A LDA(Latent Dirichlet Allocation)** model with higher Log-Likelihood and lower Perplexity is considered to be good.
- Here is our model perplexity and Log-Likelihood results,

  Perplexity: -9965645.21463

  Log=Likelihood: 2061.88393838

Multi Label Classification using One Vs Rest Classifier:

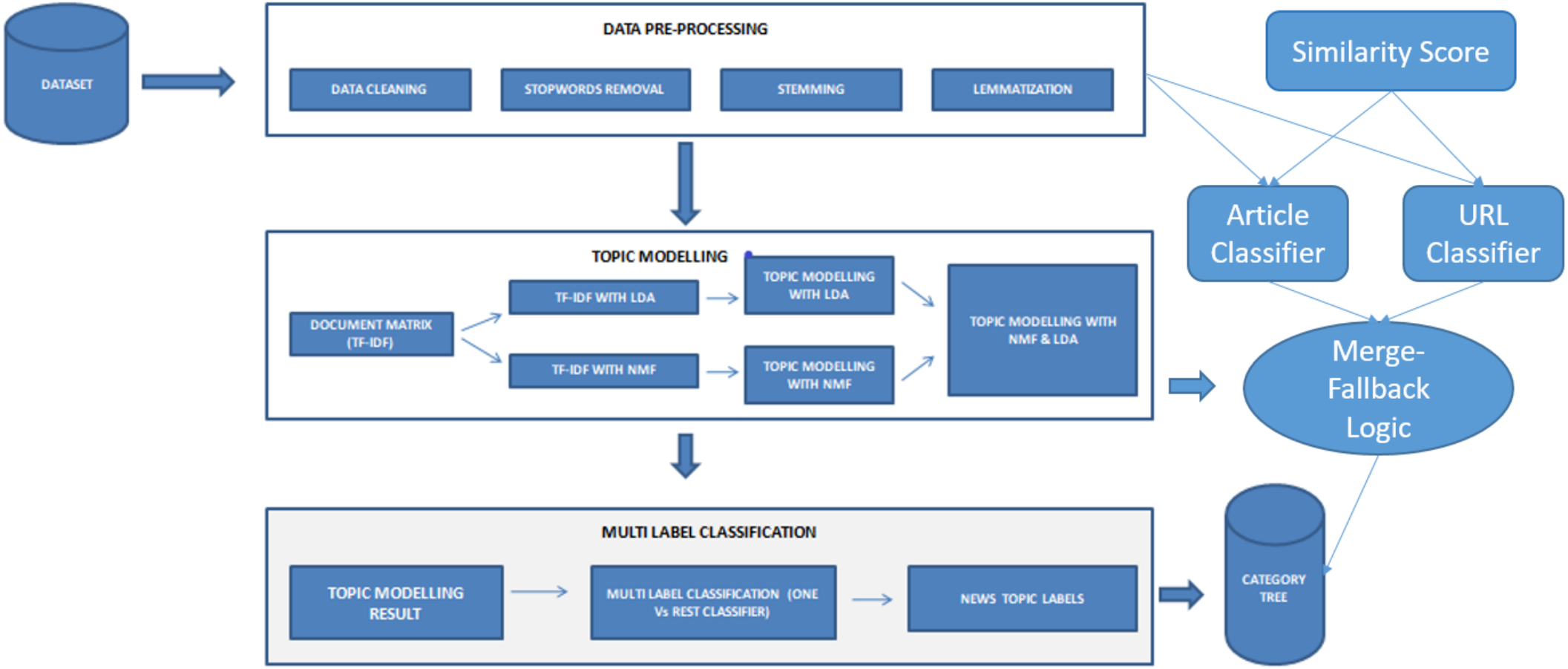
- After getting LDA and NMF outputs. I combined it and found probability of each topic assigned documents to specific topics which has highest probability.
- After that used **Multi-Class Classification** algorithm, **One Vs Rest Classifier** to split the topics,

  -> For example: consider we have keywords bjp, vote, election, modi we will assign to topic **Politics/Election**

  -> Using One Vs Rest classifier we split,
  - bjp, modi into **Politics**.
  - vote, election into **Election**.

- Evaluated **One Vs Rest classifier** model with **F1 score** Metrics and got highest score of around **81%**.

Architecture



Why our team is the best!

- I have done background reading and also put in deep thought on the given problem. There are more ideas to implement but couldn't implement due to compute and memory and human resource limitations (individual participation). I will mention some of those techniques in the next slide.

- Have in-depth knowledge in **DL, NLP, Computer Vision**.

- Have worked on all formats: **Images, Text and Speech** also as input to our Deep Learning system

- Have experience on **similar problems.**

- Moreover, there is **passion** in my work.

- I believe in **continuous improvement.**

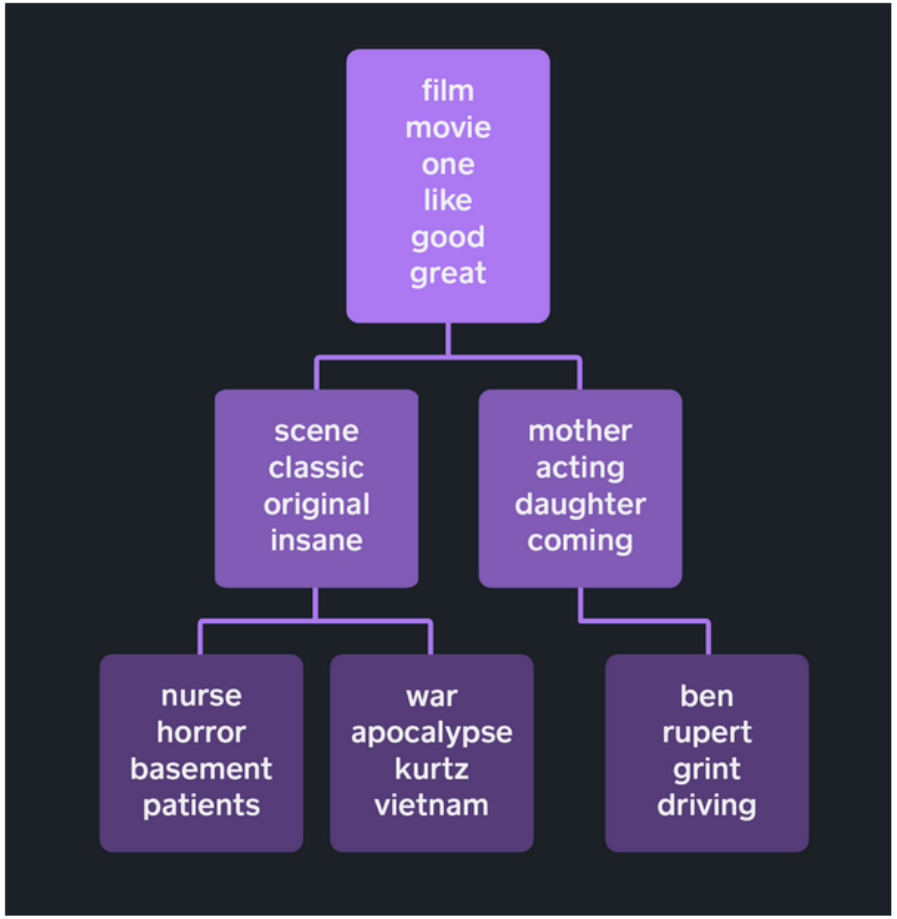
- Working as **VP of Technology** in a company which uses state-of-the-art technologies.

hLDA: Hierarchical-LDA to group documents by hierarchies of topics

```
[['movies' 'seem' 'fall' 'two' 'categories' 'films' 'reinforce'
'existing' 'societal' 'values' 'beliefs' 'challenge'],

['looked' 'movie' 'think' 'belongs' 'among' 'hitchcocks' 'greatest'
'films' 'first' 'time' 'saw' 'blown' 'away']]
```

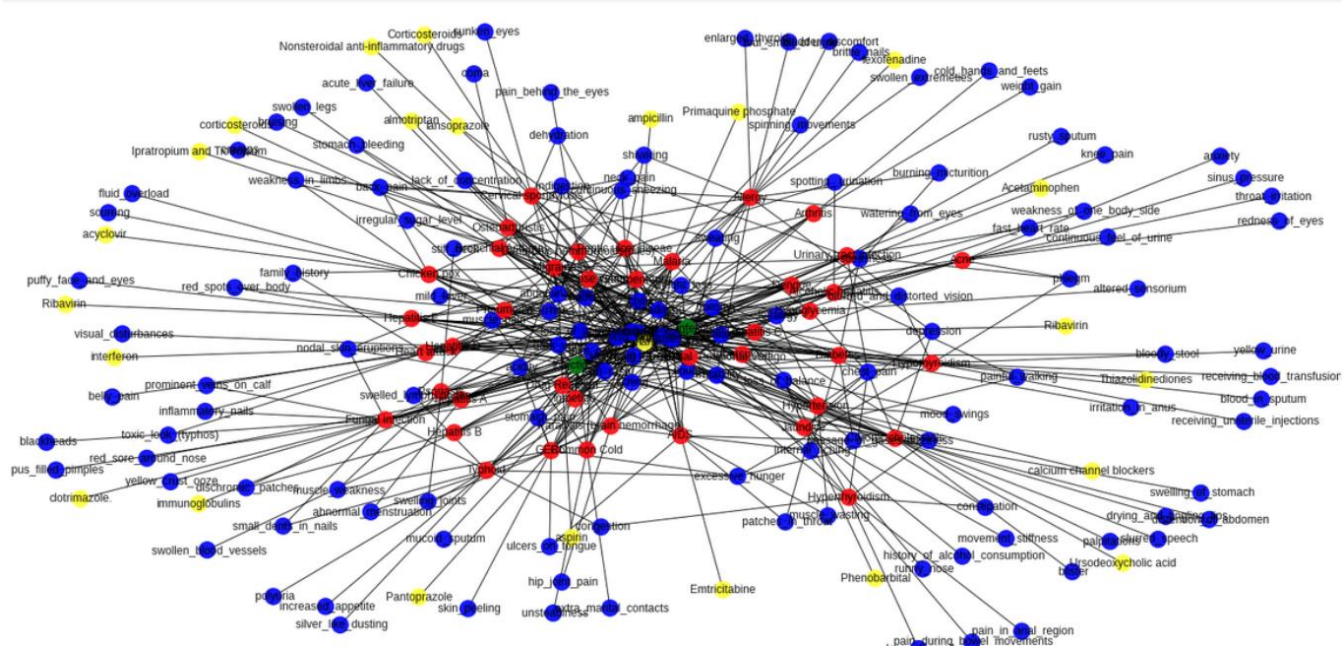
Running hLDA gave us the following (partial) results:



Improvements

Knowledge Graph

Knowledge graph of article text would generate hierarchical relation of words, akin to a category tree.



Associated attachments/ files

1. Multi-Lingual_Category_Tree_Classification.ipynb: Source code in Main Code in Jupyter notebook
2. News_Classification.ipynb: Fall-back method of LDA-NMF classification
3. Result_Submission.csv: Result in the format specified
4. Input files = [Train_data.csv, one.csv, two.csv, three.csv, four.csv, four1.csv, noun.csv, result.csv,result2_today.csv]
5. Presentation: contains architecture, solution pipeline, thoughts, techniques etc.
6. Readme.txt: step by step guidance to run the code

Challenges

- First challenge was on input data. It took time to clean the data.

- Another challenge was the usage of multiple languages in the input dataset. I have used Spacy language detector & Watson API to solve.

- The classic solution to such problem was **Latent Dirichlet Allocation (LDA)** and **Non Negative Matrix Factorization (NMF)**. But for this input document set, LDA and NMF was just giving important keywords and not topics as required.

- Since **LDA** and **NMF** only gives the keywords not topics so labelled the topics based on keywords.

- As neither LDA nor NMF was giving great results, we had to find an innovative way to combine the two methods.

- After that also some miss-classification in topics happened. To solve this problem, used Multi Label Classification method **One Vs Rest** Classifier.

- There are more ideas to implement but couldn't implement due to compute and memory and human resource limitations (individual participation).

Thank You