Annex A - Theme ‘caste’ newspaper article selection

The set of words chosen for selecting the theme ‘caste’ from the set of newspaper articles was obtained through the implementation of a Word2vec (Mikolov, T. et al 2013) word vectorization over all the newspapers dataset. The result of the process is a list similar to the one on Fig A1.

```
>>> print(word_vector_model['caste'])

[ 0.36144811  0.9394744  -0.04960836  0.12697324 -0.05321285  1.26333201
 -1.92637789 -1.7936157  -0.30662745 -4.34078264  1.49992335  1.88930798
 -1.5879575  -1.68367743  1.05037796 -0.05518804  1.06328952 -0.65698749
 -0.43360701  0.55273622 -1.95642138 -0.01125649  1.43646395 -2.50937271
 -1.36339295  0.39476418  2.60586357 -1.29740882  1.20193529  2.79373574
 -0.69485408  1.55039299 -0.28078166 -0.35374898  0.11597534 -1.03041029
  0.74071085  2.12659192 -0.17358382  0.83559793 -0.79655439  0.47574732
 -4.80029297 -0.35882443  1.79083717 -1.02862287 -0.06136697  0.76413924
  1.17297339 -1.97265792 -1.87739408  4.39395761  0.47494641 -0.20578043
  1.81348133 -0.9901408   1.69778907  1.84521914 -0.95384908  0.31571835
  0.80007201 -0.64229447 -1.17047143 -1.77952898  0.51412797 -0.42120039
  0.23139261  0.62367862  0.40113685 -1.28240991 -0.97159761  1.22404552
  1.61761415  1.27893245 -1.98591435  0.25257453 -1.9399178  -0.97328275
 -3.87632489  0.11386199 -0.06567442  1.18908536 -1.35176075  1.75691295
 -0.88603038  1.12905872 -1.28765666  2.24893332  0.97113425  1.36996317
 -0.89293981 -4.85557175 -2.35617065 -0.21033286 -0.1270193   1.88836622
  0.56042582  3.72345591  0.5098598  -5.05014992]
```

Fig A1 - Vector representation of the example word ‘caste’ from the word vector model obtained from the texts dataset. All the words in the vocabulary of the dataset have a similar vector representation.

Normal vector operation can be applied to the vector representation of some word, namely cosine similarity between vectors:

```
>>> print(word_vector_model.similarity('caste','dalit'))
0.725515860498
```

Other geometric operations, like adding or subtracting vectors, can also be applied. This approach can be validated following classic example usually cited in the Word2Vec literature, applied to the word vector model we extracted from our dataset. In this case the sum of the

vector ‘king’ with the vector ‘woman’ with the subtraction of the vector ‘men’ results in a vector for which the ten most similar word vectors are the following:

```
>>>print(word_vector_model.most_similar(positive=['king','woman'],negat
ive=['men'], topn=10))
[(u'guest', 0.49329015612602234),
 (u'teresa', 0.4903309643268585),
 (u'mother_teresa', 0.488112211227417),
 (u'born', 0.4709968566894531),
 (u'london', 0.4677670896053314),
 (u'queen', 0.4669607877731323),
 (u'new_york', 0.4655665159225464),
 (u'prince', 0.46218621730804443),
 (u'girlfriend', 0.4588596522808075),
 (u'crown_prince', 0.4569362699985504)]
```

Or, in the case of ‘cast’ and ‘dalit’:

```
>>>print(word_vector_model.most_similar(positive=['caste'],negative=['d
alit'], topn=10))

[(u'exercising', 0.4622369706630707),
 (u'grossly', 0.45106828212738037),
 (u'classification', 0.4496675729751587),
 (u'selections', 0.4446200132369995),
 (u'proposition', 0.4413527250289917),
 (u'categorisation', 0.43715623021125793),
 (u'fairness', 0.4369693994522095),
 (u'sharia', 0.4345862865447998),
 (u'competence', 0.43454012274742126),
 (u'domain', 0.43326646089553833)]
```

This result shows give us a list of words characterizing the abstract meaning of the word caste without taking into account its concrete human instantiation in a person of a particular caste as a ‘dalit’.

In our particular case, in order to obtain a broad spectrum of words, we searched for the 7 most similar words to the word ‘caste’. Then, for each of these, we searched two more times for 7 most similar words again. We tried the process with different breaths in each iteration and also with different iteration depts. After some essays, with larger breaths or depts, the theme of the

words becomes distant from the seed word. At some point, ‘discrimination’, ‘politics’ and ‘religion’ related words emerge in the process. For each first step word we recorded the similarity measure with the root word (‘caste’), and for each iteration the similarity was multiplied by the similarity of the previous words. The resulting word similarity value was ultimately related to the root word.

The final set of words was:

```
caste =['backward_classes', 'maratha_community', 'castes', 'jatavs',
'religion_caste', 'caste_creed', 'obcs', 'scheduled_tribes',
'backward_class', 'scheduled_caste', 'patidar_anamat', 'tribes',
'dalit_community', 'maratha', 'andolan_samiti', 'jats', 'obc_quota',
'demanding_reservation', 'scheduled_castes', 'obc', 'dalits',
'obc_category', 'tribe', 'patidar_community', 'marathas',
'backward_castes', 'patels', 'yadavs', 'patidar', 'irrespective_caste',
'dalit', 'tribals', 'dalits_obcs', 'upper_castes', 'demand_reservation',
'dalit_communities', 'sc_st', 'scheduled_tribe', 'backward_communities',
'caste_religion', 'adivasis', 'sub_castes', 'brahmins', 'basis_caste',
'scs_sts', 'caste_system', 'jat_community', 'extremely_backward',
'upper_caste', 'patidars', 'scst', 'mevani', 'backward', 'caste']
```

The resulting graph of words, as depicted in Fig.1, is the result of a larger graph, pruned according to a given similarity threshold and after the removing the following religion related words:

```
['hinduism','inciting','secular', 'hindus', 'culturally', 'communities',
'religions', 'minority_communities', 'minority_community', 'creed',
'ideals', 'preaching', 'propagating', 'tolerance', 'belief', 'muslim',
'vested_interests', 'socially_educationally', 'secularism', 'minorities',
'ideology', 'muslims', 'hind', 'hatred', 'community', 'liberal',
'hindutva', 'sikh_community','hindu_muslim','hindus_muslims', 'societal',
'atrocities', 'values', 'beliefs', 'muslim_community', 'communalism',
'divisive', 'nationalism', 'portray', 'support_base', 'seats_reserved',
'ideologies', 'minority', 'fascist', 'philosophy', 'buddhists',
'propagate', 'christians', 'intellectual', 'muslims_christians',
'particular_community', 'ideological']
```

Having defined the set of words potentially associated with the theme caste, a measure of the density of these words in each article was calculated. This measure was obtained multiplying the

similarity of each of the selected words in the ‘caste’ set by its count in each article, and dividing by the total number of words in the article.

The cumulative probability of an article having theme ‘caste’ related word density larger than a certain value in our dataset is depicted in Fig.A2.

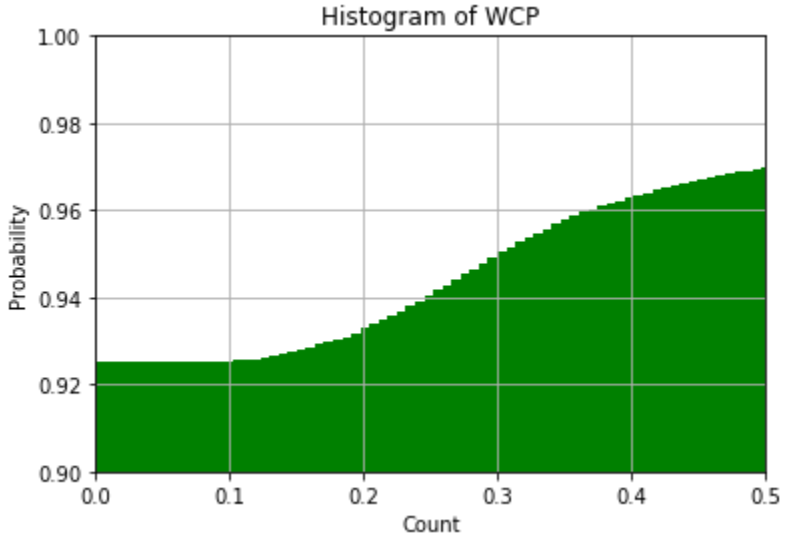

Fig.A2: Cumulative probability of an article having theme ‘caste’ related word density from the whole dataset.

From this result, we choose a threshold density of 0.15, from which around 7% of the articles were selected. These articles constituted our ‘caste’ related sub dataset.

Annex B - Examples of topics

In the following two tables some detected topics, first from the subset ‘caste’ and in second remaining articles, are reported. At the left of each word the corresponding probability of occurrence within the topic.

Social and political violence		**Reservation demand**		**Student politics**		**Status of SC/ST/OBC**		**Religion**	
0.019	**army**	0.047	**patel**	0.024	**students**	0.013	**bill**	0.013	**kovind**
0.017	**terror**	0.023	**patidar**	0.014	**kerala**	0.011	**commission**	0.010	**hindu**
0.015	**security**	0.018	**hardik**	0.012	**university**	0.010	**act**	0.010	**muslims**
0.011	**injured**	0.013	**rupani**	0.010	**student**	0.009	**committee**	0.009	**muslim**
0.010	**border**	0.011	**surat**	0.008	**suicide**	0.009	**report**	0.009	**religion**
0.009	**fire**	0.010	**ahmedabad**	0.007	**college**	0.009	**scheduled**	0.007	**society**
0.008	**lynching**	0.009	**agitation**	0.006	**rss**	0.007	**tribal**	0.006	**rss**
0.007	**violence**	0.009	**rally**	0.006	**karnataka**	0.007	**justice**	0.006	**right**
0.007	**militants**	0.008	**patidars**	0.006	**education**	0.007	**land**	0.006	**hindus**
0.007	**beats**	0.007	**sardar**	0.005	**hyderabad**	0.006	**ministry**	0.005	**religious**
0.007	**crpf**	0.007	**yatra**	0.005	**slaughter**	0.006	**rights**	0.004	**name**
0.006	**forces**	0.007	**event**	0.005	**school**	0.005	**status**	0.004	**please**
0.006	**maoists**	0.007	**nitin**	0.005	**campus**	0.005	**backward**	0.004	**nation**
0.005	**personnel**	0.006	**protest**	0.004	**jnu**	0.005	**order**	0.004	**say**
0.005	**mob**	0.006	**youths**	0.004	**union**	0.004	**issued**	0.004	**think**
0.005	**leaked**	0.006	**members**	0.004	**death**	0.004	**department**	0.004	**human**
0.005	**trump**	0.006	**samiti**	0.004	**protest**	0.004	**passed**	0.004	**love**
0.005	**area**	0.006	**vijay**	0.004	**ryan**	0.004	**tribes**	0.004	**violence**
0.005	**peace**	0.006	**quota**	0.004	**telangana**	0.004	**secretary**	0.003	**never**
0.005	**towards**	0.005	**mevani**	0.004	**doctorate**	0.004	**panel**	0.003	**become**
Coherence measure:									
0.449		0.725		0.385		0.516		0.453	

Politics		Celebrity		Insurgency		Women specific crime (dowry, infanticide)		Cricket	
0.028	**bjp**	0.039	**sharma**	0.021	**kashmir**	0.011	**school**	0.017	**vs**
0.020	**congress**	0.031	**kapil**	0.019	**army**	0.010	**old**	0.011	**twitterati**
0.020	**party**	0.025	**rai**	0.017	**pakistan**	0.009	**year**	0.010	**refugees**
0.013	**chief**	0.021	**delhi**	0.014	**security**	0.008	**murder**	0.010	**salman**
0.010	**leader**	0.018	**bachchan**	0.010	**border**	0.008	**woman**	0.010	**score**
0.008	**state**	0.017	**kejriwal**	0.010	**troops**	0.008	**family**	0.010	**read**
0.008	**modi**	0.014	**aishwarya**	0.010	**jammu**	0.007	**daughter**	0.009	**cricket**
0.008	**singh**	0.014	**karnataka**	0.010	**killed**	0.006	**wife**	0.009	**odi**
0.007	**leaders**	0.013	**ranbir**	0.010	**forces**	0.006	**student**	0.009	**australia**
0.007	**president**	0.012	**dual**	0.009	**attack**	0.006	**girl**	0.009	**unga**
0.007	**election**	0.011	**goa**	0.008	**terror**	0.006	**man**	0.008	**live**
0.007	**yadav**	0.011	**show**	0.008	**militants**	0.006	**father**	0.008	**world**
0.007	**nitish**	0.009	**sep**	0.008	**ceasefire**	0.005	**us**	0.008	**judwaa**
0.007	**kumar**	0.009	**impressions**	0.007	**valley**	0.005	**one**	0.007	**nadal**
0.006	**assembly**	0.009	**arvind**	0.006	**kusa**	0.005	**read**	0.007	**varun**
0.006	**gandhi**	0.008	**believable**	0.006	**situation**	0.005	**case**	0.007	**jacqueline**
0.006	**elections**	0.008	**free**	0.006	**along**	0.005	**police**	0.007	**data**
0.005	**political**	0.008	**plus**	0.006	**home**	0.005	**women**	0.007	**expressgroup**
0.005	**cm**	0.008	**water**	0.005	**pakistani**	0.004	**students**	0.006	**photo**
0.005	**kovind**	0.008	**tells**	0.005	**operation**	0.004	**son**	0.006	**deol**
Coherence measure:									
	0.520		0.518		0.474		0.399		0.447