PRINT MEDIA AND AADHAAR

Text preprocessing | Topic mining | Clustering | Sentiment analysis

APRIL 16, 2018

ANIRUDH SYAL | KRITI BHOLA | SV RAVI CHANDRA
Indian school of business

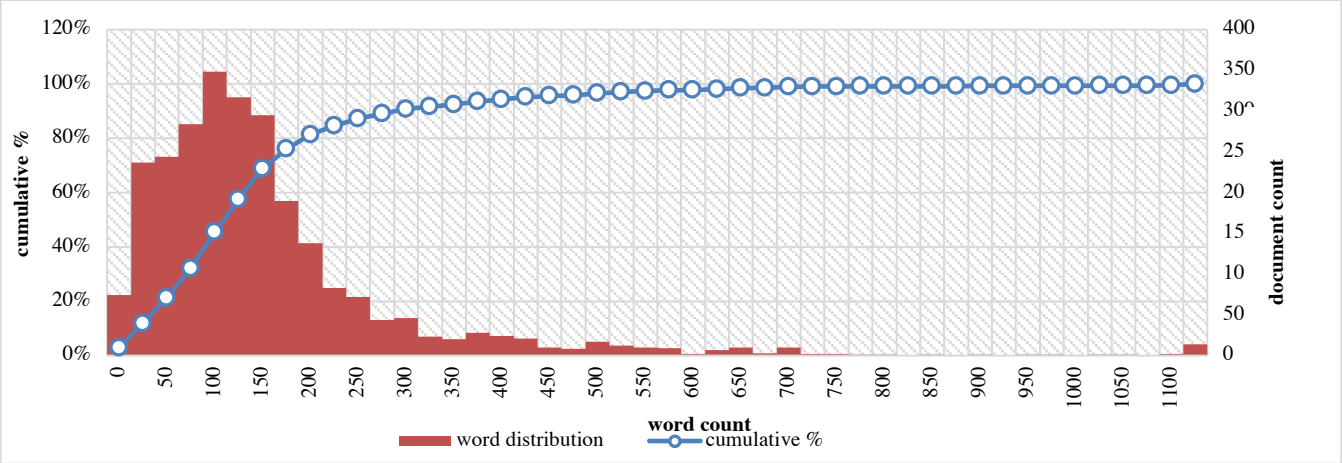
Print Media and aadhaar

This data brief is an exploratory analysis of the various topics covered by print media around the core theme of 'aadhaar'. The scope of this work is limited to articles and archives of The Hindu. We intend to discover various topics, their emergence with time, asses the case of scope creep in 'aadhaar' and capture the sentiment polarity attached with these topics. We have scraped over **3000** news articles containing the term 'aadhaar' and its derivatives, from the archives of Hindu. Below, we describe the initial results of this activity.

Data Exploration / Descriptive analysis

Plot of observations

Throughout this data brief, we have performed all analysis after the elimination of stop words / non informative words. **Figure 1** is the plot of the distribution of aadhaar centric articles by words in each quarter from 1st quarter of 2006 to 1st quarter of 2018. **Figure 2** is the plot of distribution articles by the number of times the term aadhaar appears in the article. Throughout this data brief, we refer to the 1st quarter as the first three calendar months (Jan-Mar). The average article length is 180 words, while the median article length is 136 words. In this data brief, we use the terms 'token' and 'word' interchangeably. The term aadhaar appears exactly once in 50% of the articles.



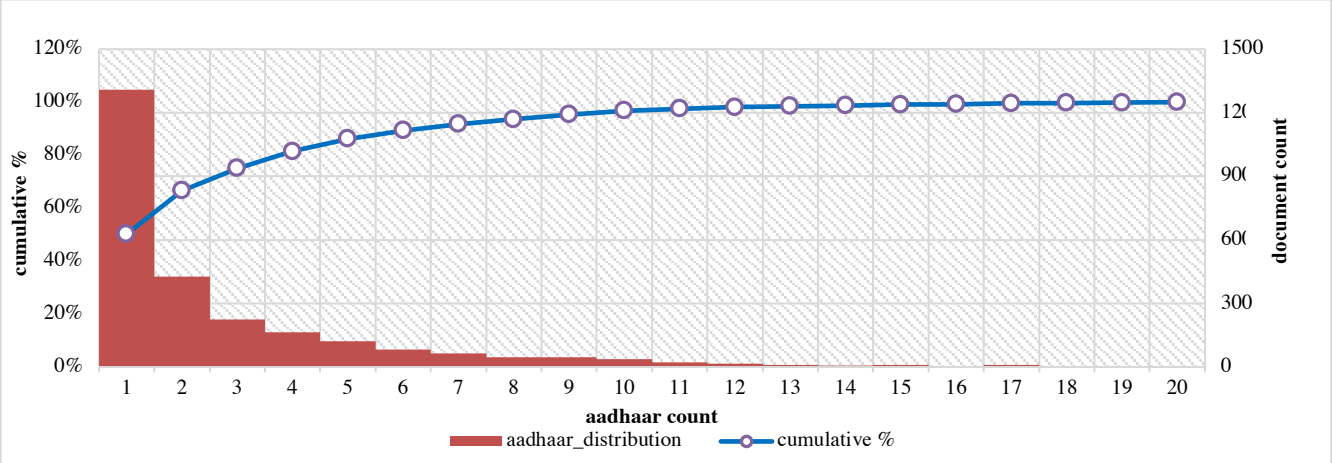
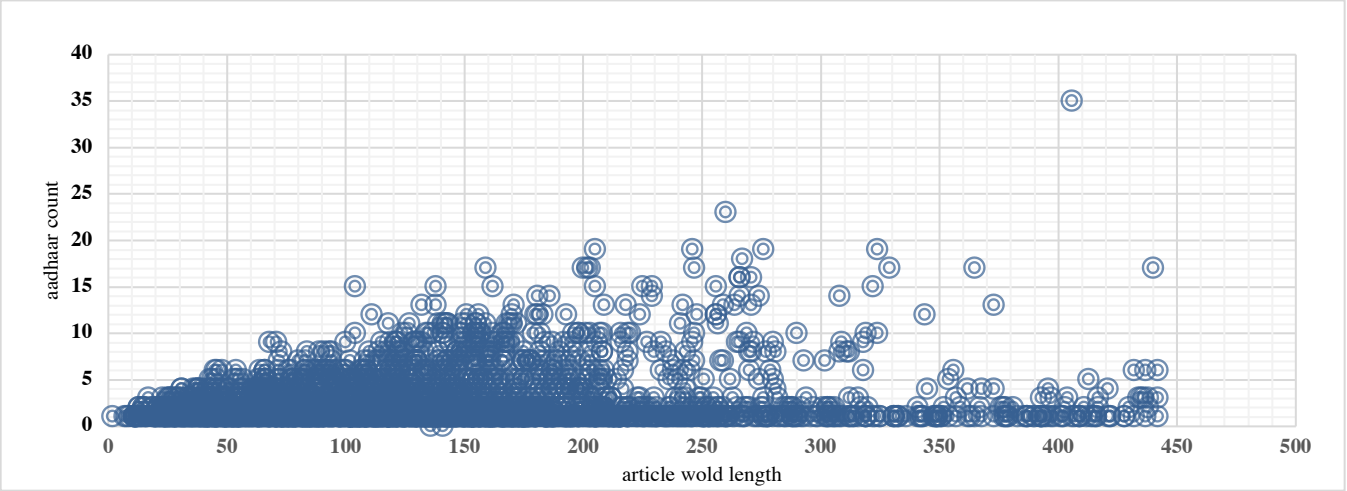
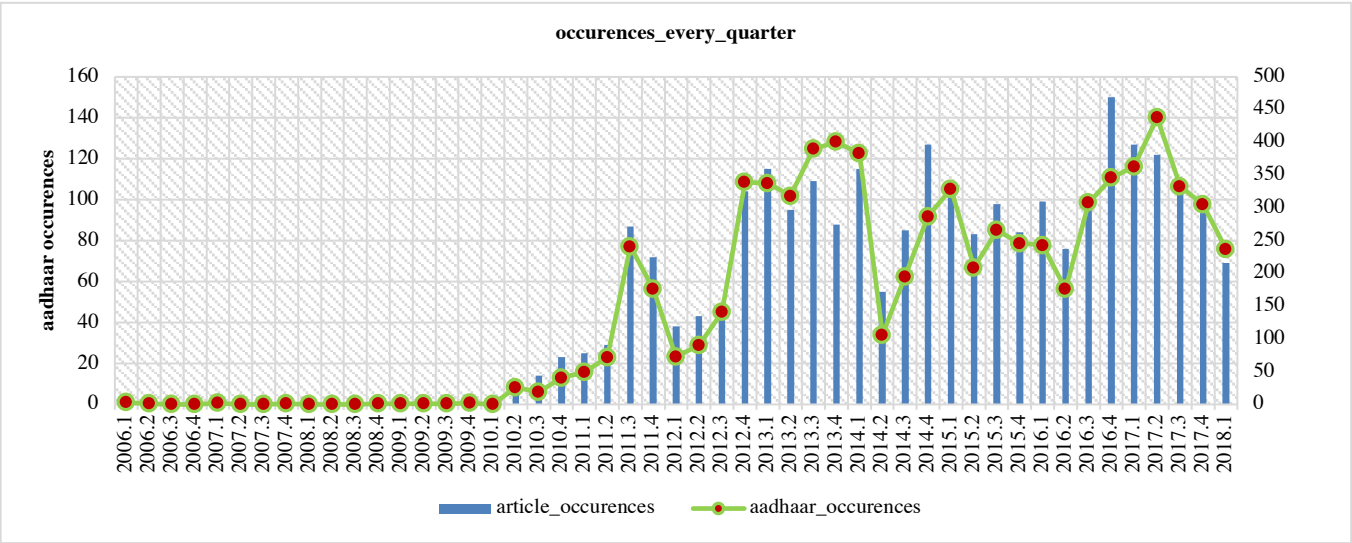


Figure 3 is a scatter plot of count of occurrences of aadhaar in an article versus the article's word length.

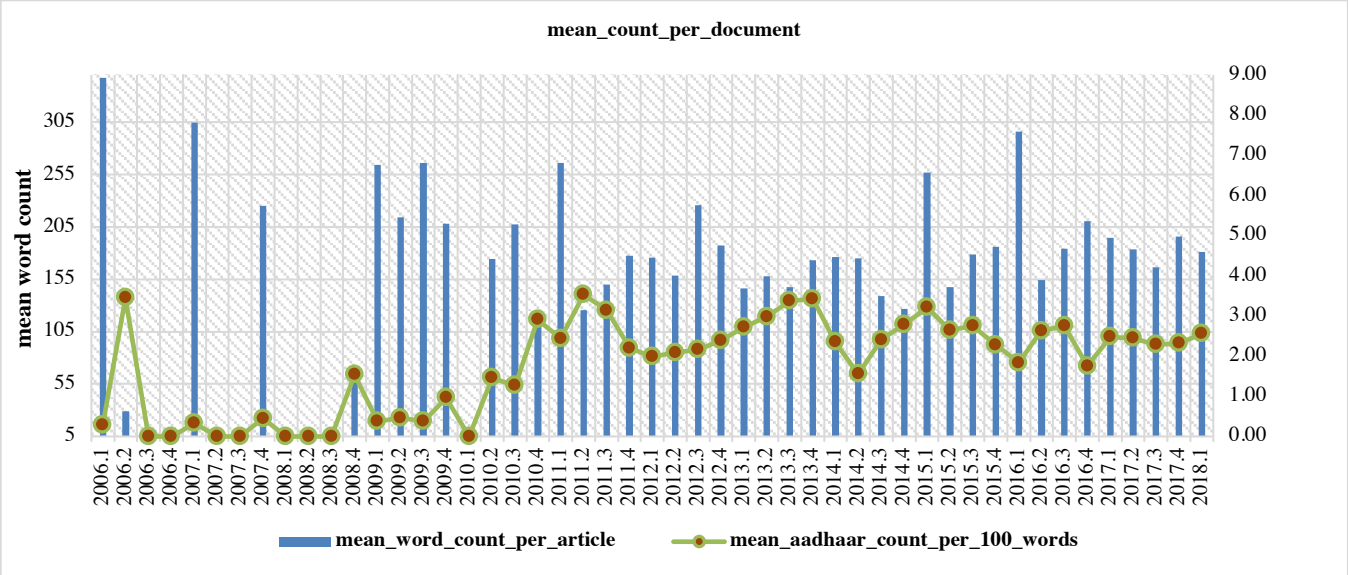


Plots grouped by quarter

Figure 4 , is a time series plot of the emergence of aadhaar centric discourse in print media over 49 quarters (proxied by number of articles containing the term 'aadhaar'). The **primary 'y' axis** represents the total occurrences of the term 'aadhaar' in a given quarter . The **secondary 'y' axis** represents the count of articles. Clearly, the discourse around aadhaar has gained momentum since 2010 and is a widely discussed subject matter, with over 300 articles every 3 months, implying over 3 articles each day that contain the term 'aadhaar'.

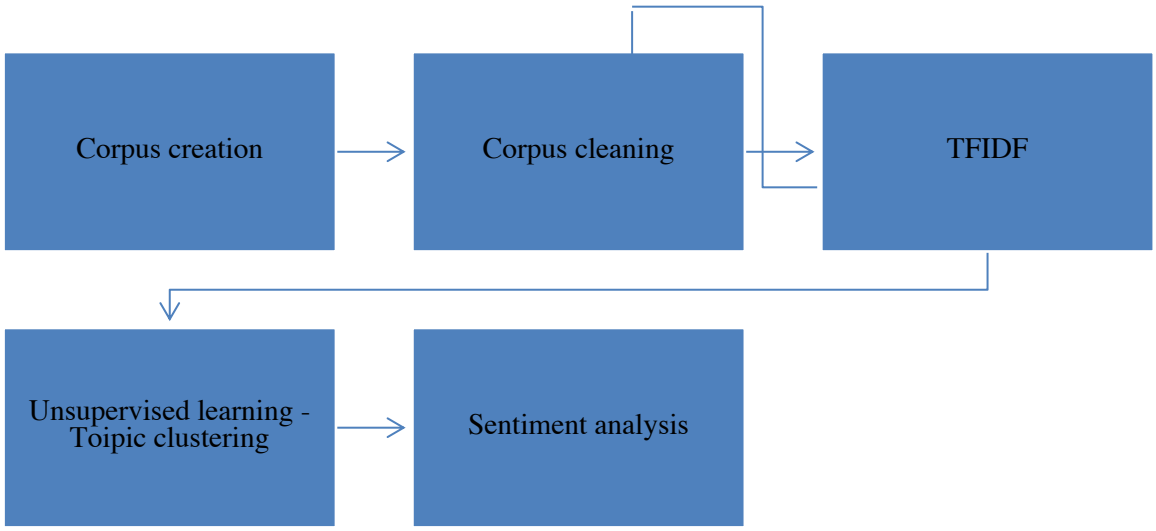


In **figure 5** , we plot for each quarter, the mean word count per an aadhaar centric article on the primary 'y' axis and the mean 'aadhaar' occurrences per 100 article words on the secondary 'y' axis.



Data Preparation

General Methodology



Corpus Creation

The corpus was collected by scrapping live news website of Hindu and its archives.

- News articles from all three website were searched for keyword “Aadhaar” from archives dating from 2006 to 2018.
- Selenium web driver was used to scrape content as the host website had dynamic content.
- Beautiful Soup and requests libraries were used to extract the news heading, date, and news body.
- The extracted corpus was pickled in the form of data frames to be used by a python program.

Corpus Preparation

The corpus was pre-processed to remove HTML junk, stop-words, unwanted digits, punctuation characters, etc. The initial cleaning was done in 4 stages.

Stage 1 and Stage 2

- Eliminate HTML junk
- Replace all text as would’ve by would have etc.
- Eliminate dangling or repeated non alpha numeric content such as ‘….’
- Match regex patterns to combine high frequency bi grams and non- alpha numeric text such as percentage signs, dollar signs etc.

Stage 3

- Eliminate stop words
- Lemmatize words using **Part of Speech tagging**
- Eliminate dangling digits and clean more junk at the scale of each word

Stage 4

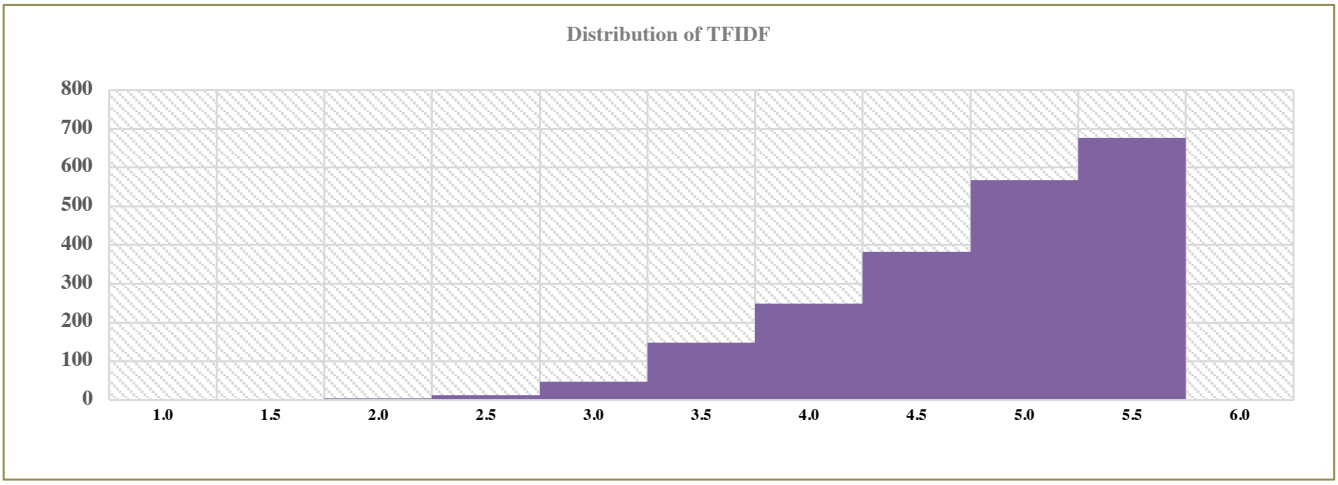
- Replace adverbs and adjectives by the root verb

Stage 5

- Generate term frequency inverse document frequency (TFIDF) matrix
- Assess high frequency bigrams and trigrams in the corpus
- Replace the high frequency co-occurring unigrams with the respective bi grams and tri grams
- Re-Generate TFIDF

TFIDF ANALYSIS AND WORD CLOUD

The plot below is a histogram of the distribution of words with respect to their TFIDF. This is then mapped to form a word cloud showcasing top 1000 words as they appear in the corpus, the font size being relative to the tfidf.



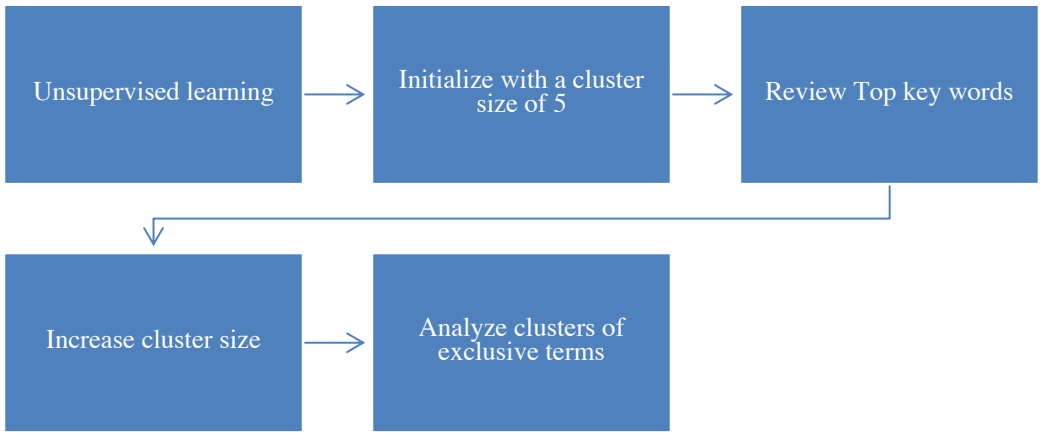
TFIDF mapped to a word-cloud

Terms	TFIDF
aadhaar_linkage	5.4
aadhaar_target_delivery _financial_subsidy	5.4
andhra_pradesh_ government	5.4
avail_subsidy	5.4
civil_supply_minister	5.4

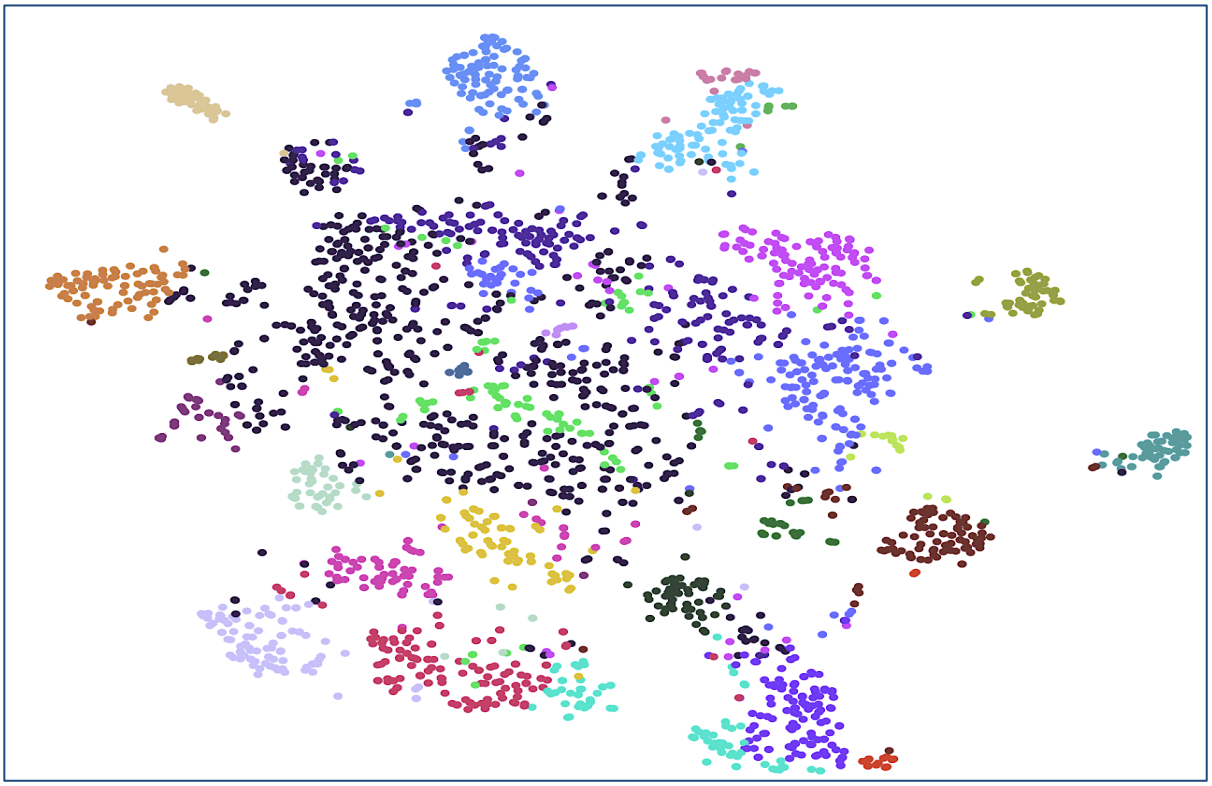
UNSUPERVISED CLUSTERING

K nearest neighbors

In order to cluster the documents , the first technique that was used was the K-NN clustering technique.
In order the visualize the clusters, the corpus of documents was reduced to a vector space of two dimensions using the t-distributed stochastic neighbor embedding technique (T-SNE). The T-SNE using 2 -components reduces the dimensionality of a corpus of documents in the n -feature space to two dimensions. This allows each document to be plotted in 2-D space for visualization using 'bookeh plots'



For each cluster, we analyzed the **top 20 highest information terms** of the cluster to decide whether further clustering was required. Below, is a 'bookeh' plot representing 29 clusters in **2-D space.**



S.no	Cluster Terms
1	police | arrest | document | accuse | aadhaar_card | case | fake | victim | crime | woman | alleged | house | incident | police_station | complaint | person | passport | city | investigation | duo |
2	card | ration_card | smart_card | aadhaar_card | white | bogus | survey | apply | district | link | ration | centre | fair_price_shop | government | department | office | supply | issue | shop | aadhaar |
3	npr | resident | direct_census | camp | operation | national_population_register | data | chennai | census | biometric_data | process | biometric | collect | data_collection | corporation | district | enrolment | aadhaar_number | detail | form |
4	housing | house | slum | rent | land | scheme | construction | sanction | group | government | project | beneficiary | colony | benefit | demand | loan | family | urban | reservation | subsidy |
5	centre | aadhaar_card | road | vehicle | card | registration | applicant | post_office | service | bus | office | city | district | application | detail | number | issue | register | resident | traffic |
6	budget | india | government | tax | sector | growth | economy | fiscal | country | economic | increase | reform | scheme | revenue | state | investment | policy | indian | development | gdp |
7	student | college | school | education | department | scholarship | camp | university | principal | art | class | aadhaar | child | teacher | government | engineering | state | engineering_college | aadhaar_card | institute |
8	ward | resident | camp | municipal | corporation | registration | colony | office | nagar | aadhaar_card | biometric_data | photograph | register | family | school | high_secondary_school | centre | visit | member | issue |
9	system | ration | minister | babu | distribution | district | state | aadhaar-enabled | chief_minister | public_distribution | fair_price_shop | ration_shop | cashless_transaction | programme | cardholder | shop | team | government | introduce | point_sale |
10	child | hospital | school | health | parent | work | job | woman | baby | patient | foundation | help | education | family | aadhaar_card | government | social | doctor | home | village |
11	vendor | street | token | association | grievance | food | system | resident | bill | woman | cashless_transaction | vegetable | business | authority | detail | lack | member | livelihood | locality | address |
12	pan | mandatory | aadhaar | income_tax | file | apply | petition | court | act | aadhaar_number | pan_card | student | scholarship | government | supreme_court | return | link | centre | section | tax |
13	district | post_office | customer | aadhaar_card | registration | collector | account | worker | joint_collector | counter | panchayat | camp | submit | link | bank_account | seed | complete | release | scheme | special |
14	consumer | lpg | bank_account | link | cylinder | subsidy | district | aadhaar_number | direct_benefit_transfer | distributor | bank | refill | oil_company | scheme | petroleum | lpg_subsidy | customer | aadhaar_card | subsidy_amount | aadhaar |
15	digital | bank | customer | payment | cable | company | transaction | business | village | atm | india | mobile | product | service | tv | wallet | finance | cash | country | technology |
16	bill | money_bill | rajya_sabha | lok_sabha | amendment | aadhaar_bill | pass | opposition | house | government | parliament | supreme_court | speaker | service | benefit | congress | aadhaar_target_delivery_financial_subsidy | india | law | aadhaar |
17	election | ghmc | commissioner | election_commission | electoral | letter | hyderabad | voter | seed | chief | epic | implementation | poll | visit | door-to-door | officer | electoral_roll | detail | complaint | announcement |
18	bench | aadhaar_card | justice | interim_order | order | petitioner | court | supply | petition | petroleum | file | supreme_court | lpg_cylinder | high_court | insist | person | hearing | mandatory | authority | apex_court |
19	bank | account | branch | customer | bank_account | banking | deposit | district | payment | note | cash | facility | currency | money | open | manager | beneficiary | scheme | launch | rbi |
20	airport | recruitment | test | passenger | candidate | physical | security | international | ceo | entry | technical | hyderabad | conduct | body | personnel | second | strength | post | apply | admit |
21	voter | electoral_roll | election | voter_list | epic | vote | booth | polling | constituency | polling_station | officer | election_commission | assembly | district | voting | photo | slip | link | deletion | delete |
22	aadhaar | india | uidai | data | indian | government | country | project | state | identity | work | technology | world | issue | national | like | service | system | information | film |
23	scheme | government | party | cpi | beneficiary | leader | subsidy | congress | aadhaar_card | delhi | price | state | implement | direct_benefit_transfer | cash_transfer_scheme | demand | supply | centre | chief_minister | link |
24	pension | pensioner | passport | certificate | employee | employer | document | beneficiary | submit | aadhaar_number | office | issue | pension_scheme | application | government | organisation | employee_provident_fund | officer | detail | social_security |
25	farmer | crop | loan | agriculture | loan_waiver | debt | scheme | government | land | bank | account | agricultural | district | state | farm | subsidy | amount | centre | cooperative | avail |
26	privacy | right_privacy | supreme_court | bench | fundamental_right | court | aadhaar | right | constitution_bench | government | social_medium | constitution | justice | personal_information | aadhaar_scheme | petition | voluntary | law | judgment | argue |
27	enrolment | uidai | enrol | district | cover | population | akshaya | aadhaar_card | aadhaar | centre | agency | unique_identification_authority_india | state | issue | aadhaar_enrolment | enrolment_process | process | resident | direct | person |
28	bill | price | foodgrains | household | food_security | food | family | beneficiary | scheme | bpl | state | government | tonne | provide | wheat | public_distribution | minister | system | public | subsidise |
29	railway | bsnl | ticket | customer | district | train | passenger | mobile | plan | system | manager | kerala | user | introduce | service | sim_card | office | department | state | akshaya |

UNSUPERVISED TOPIC EXPLORATION

Latent dirichlet allocation

Again, in order to explore topics iteratively, we used the LDA algorithm to determine the relevant, mostly mutually exclusive topics that could be discovered within the corpus.

LDA with 5 topics:

INFERRED TOPIC	LDA GENERATED TOPIC
General - beneficiaries of aadhaar- education / health / child/ women enrolment /	student child school police road work woman city college family education house area railway district village bus hospital health person
Direct_benefit_Transfer (LPG / Subsidy / gas /petroleum)	scheme district beneficiary consumer bank state bank_account link subsidy account lpg farmer pension aadhaar_number system minister direct centre cylinder state_government
Judiciary , aadhaar act , aadhaar bill	bill state issue supreme_court party court right congress india case order act law mandatory person public question election amendment privacy
Aadhaar_enrolmen_process and documentation	centre card number detail issue district document enrolment office resident data uidai process aadhaar_number registration voter department register service online
Technology reform + digitization + business	india country bank service indian sector like tax provide well increase system technology work project company world high digital help

LDA with 10 topics:

INFERRED TOPIC	LDA GENERATED TOPIC
General - confounded with culture, social media, india	student school child college education department camp health state conduct programme teacher class district special hospital university job medical patient
Aadhaar_enrolment_documentation, Aadhaar_enrolment_process	centre card detail enrolment issue district number office process registration document resident uidai data aadhaar_number application register passport enrol submit
Aadhaar_enabled_digital_banking/ cashless transaction	bank customer digital account service payment transaction banking user system mobile cash note branch money atm launch company mobile_phone facility
General - beneficiaries of aadhaar- education / health / child/ women enrolment /	road city area railway bus vehicle house work train nagar water family authority resident issue colony due visit police traffic
Politics, elections, aadhaar seeding, lpg subsidy	consumer lpg link voter district subsidy bank_account cylinder aadhaar_number election bank price supply customer electoral_roll gas number seed petroleum state
Aadhaar_relatd_Crime , politics	police delhi party case congress woman award chief_minister arrest national india election state leader candidate mumbai president accuse assembly person
Technology reform, aadhaar related data	india work like country world right indian well social project big technology good data part become change problem help poor
Judiciary, petition, right to privacy	bill supreme_court mandatory issue court order privacy right scheme act state benefit law bench amendment person authority concern public petition
Macro-economy, Tax, GST	india sector tax country budget growth increase provide economy fund service propose high reform policy act development infrastructure rate economic
Pension benefit transfer , food subsidy, loan	scheme district beneficiary state farmer pension minister system benefit worker state_government account family bank_account programme ensure food household subsidy loan

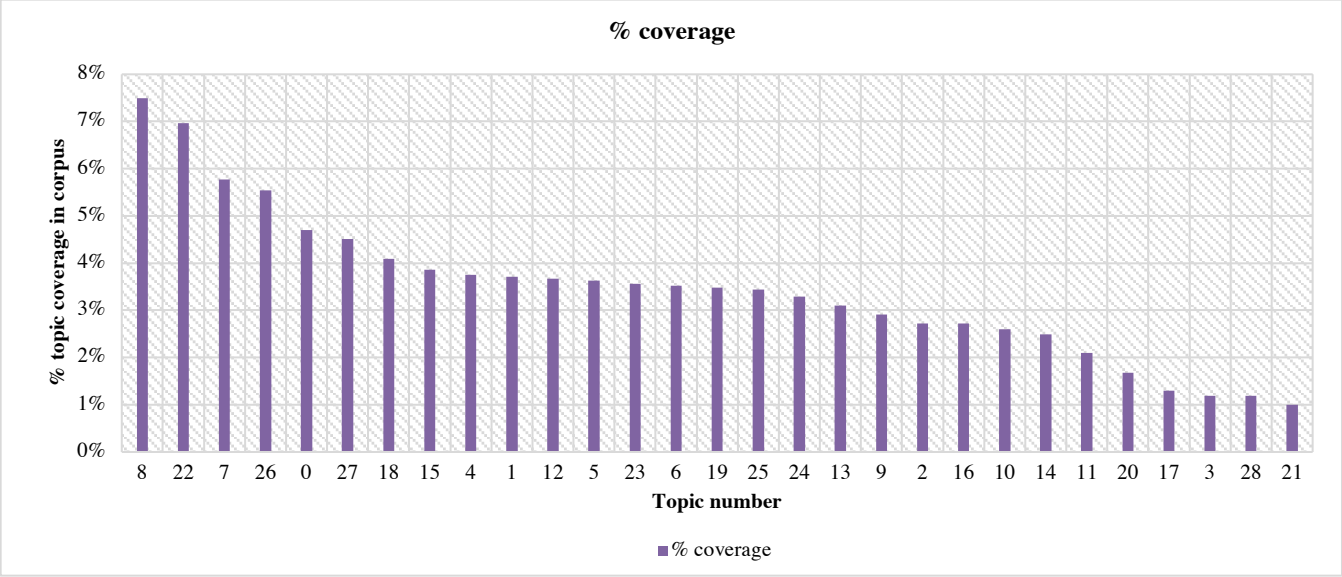
LDA with 15 topics:

INFERRED TOPIC	LDA GENERATED TOPIC
aadhaar_enabled_digital_banking	bank account payment customer banking cash transaction note branch bank_account money atm deposit rbi currency credit exchange machine open banker
aadhaar_enabled_digital_platforms , personal_data	service digital data technology system company user india internet mobile online access information launch offer plan provide platform work country
general - education / health / child/ women enrolment / beneficiary of aadhaar	student school child education health parent teacher scholarship college class patient job candidate help hospital university conduct medical doctor test
aadhaar_related_crime , Voter_id, Link with aadhaar	police voter election case document officer arrest vote district electoral_roll person constituency booth card epic work survey complaint found conduct
aadhaar_enrolment_documentation + aadhaar_enrolment_process	detail document centre submit registration passport certificate application issue register number card form online apply applicant aadhaar_number office process department
PDS, Food_security , MNREGA (Rural employment), farmer_subsidy	scheme beneficiary farmer state worker benefit subsidy provide family food household poor loan system nrega pension state_government bank_account implement account
politics - protest	delhi party congress state india chief_minister minister leader award prime_minister president election centre national present modi gujarat protest launch woman
direct_benefit_Transfer (LPG / Subsidy / gas /petroleum)	consumer link lpg bank_account subsidy cylinder aadhaar_number district scheme direct_benefit_transfer bank customer supply price gas petroleum oil_company gas_agency lpg_cylinder state
general - confounded with culture, social media, india	india like work indian country right well world social become problem change good big back part money different think report
aadhaar_database_uidai_biometric_unique_identity + aadhaar_enrolment process	uidai enrolment resident data enrol centre aadhaar_number card project district state uid unique_identification_authority_india process direct number biometric npr agency issue
general	camp college woman programme special member department film organise ward association school district national sport state office final corporation group
dbt_pension_aadhaar	district centre issue state card department meeting pension complete system programme direct ration_card collector link officer office seed public minister
aadhaar_act_parliament / right to privacy / supreme court ruling / fundamental / money bill	bill supreme_court issue court order mandatory privacy law right person act bench public petition amendment case question file rajya_sabha justice
general_infrastructure _waster / road / electricity / garbage / slum/ vehicle	road city bus area railway train nagar authority resident water colony work traffic near vehicle due street house board passenger
Macroeconomy _ tax_GST , investor, blackmoney , business, asset	india sector tax budget growth increase country economy fund provide policy development act propose infrastructure economic high investment rate market

Similarly, **LDA was done for up to 40 topics** and the team agreed that 29 topics succinctly described the entire corpus. All LDA visualizations in PYLDAVIS have been included with the report.

LDA with 29 topics:

Topic number	Main inferred topic	Sub Topic
0	**aadhaar_related_crime**	fake_card | misuse
1	**link_aadhaar_scholarships**	aadhaar_enabled_scholarships | exam_participation
2	**link_aadhaar_bank_accounts**	Financial_Inclusion | Jan_dhan_accounts
3	**aadhaar_as_document_proof**	Vehicle | driving license |liquor permit
4	**general**	transforming_india
5	**aadhaar_public_distribution_system**	food_security | aadhaar
6	**link_aadhaar_voting**	voting | voter_id
7	**aadhaar_enrolment**	aadhaar_authentication_databse
8	**aadhaar_implementation**	aadhaar_pilot_testing
9	**aadhaar_database**	data security _ efficiency _ public | pvt _sector_role
10	**aadhaar_implementation**	aadhaar_related_operational_concerns
11	**general**	healthcare_initiatives | rehabilitation
12	**aadhaar_enrolment**	aadhaar_related_documentation
13	**aadhaar_direct_benefit_transfer**	crop_loan_waiver | subsidy
14	**macro_economic Reform**	manufacturing | real Estate | GST | Banking
15	**aadhaar_legal_validity**	right_to_privacy | linking _dbt_to_aadhaar
16	**aadhaar_direct_benefit_transfer**	sarva_siksha_abhyaan
17	**aadhaar_bill**	parliamentary_process | recommendation | aadhaar bill
18	**general**	general _infrastructure
19	**link_aadhaar_to_other_documents**	aadhaar_made_mandatory_Pan_card_PF_mobile_number
20	**aadhaar_stack_digital_banking**	benefits | demonitization |digital_payment| cashless_economy
21	**aadhaar_related_debates**	Adhaar_centric_debates_social_media
22	**aadhaar_enrolment**	aadhaar_card_application | aadhaar_card_update
23	**politics**	political_protest
24	**aadhaar_stack_digital_india**	digital_platforms | internet |software
25	**general**	rural_tribal_society
26	**aadhaar_direct_benefit_transfer**	LPG _subsidy
27	**aadhaar_direct_benefit_transfer**	rural_employment | subsidy | scholarship
28	**general**	politics | culture | achievement



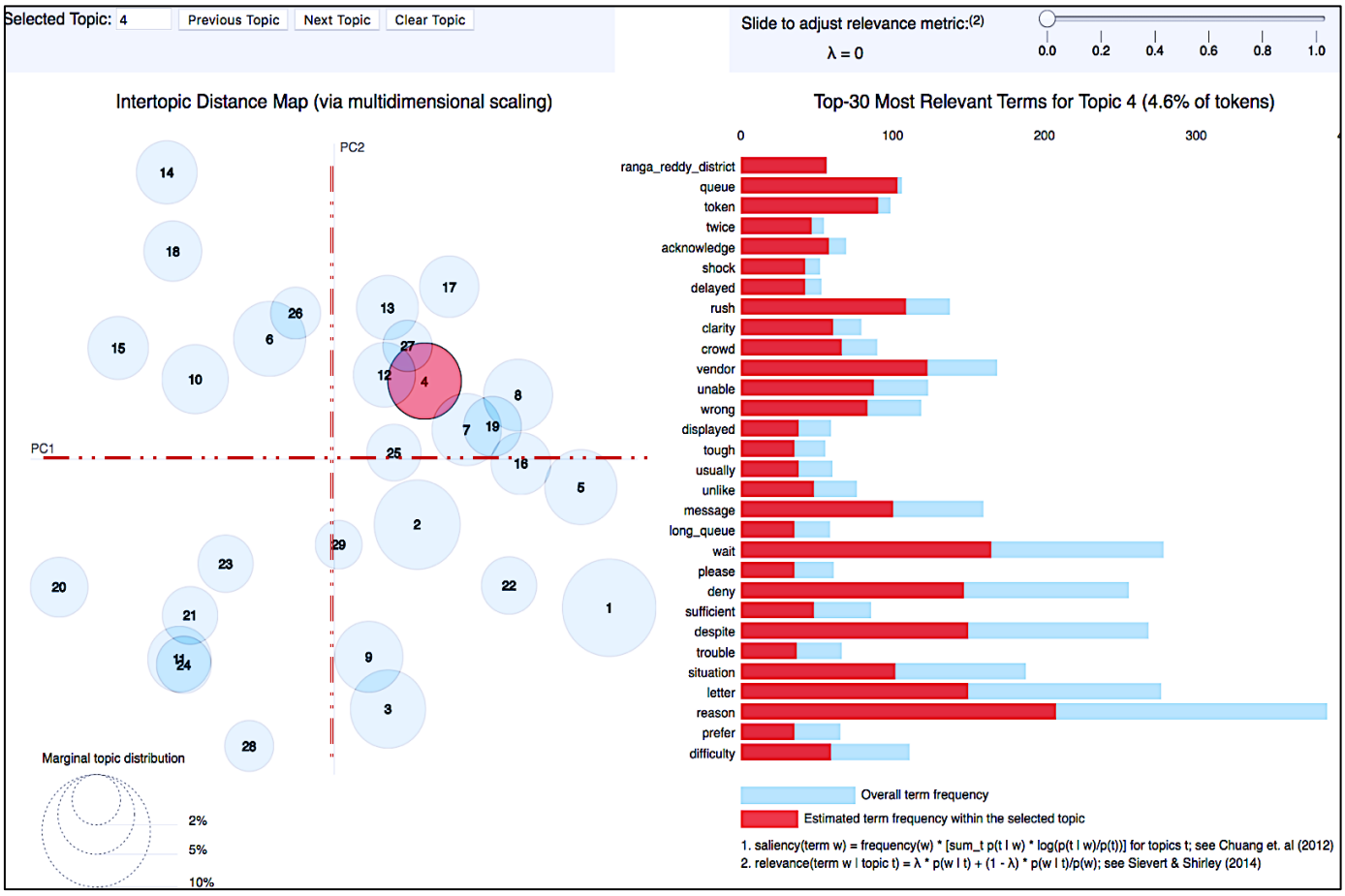
The following 14 topics cover over 65% of text corpus

	Main inferred Topic	sub topic	% coverage	cumulative %
1	aadhaar_implementation	aadhaar_pilot_testing	7%	7%
2	aadhaar_enrolment	aadhaar_card_application | aadhaar_card_update	7%	14%
3	aadhaar_enrolment	aadhaar_authentication_databse	6%	20%
4	aadhaar_direct_benefit_transfer	LPG _subsidy	6%	26%
5	aadhaar_related_crime	fake_card | misuse	5%	30%
6	aadhaar_direct_benefit_transfer	rural_employment | subsidy | scholarship	5%	35%
7	general	general _infrastructure	4%	39%
8	aadhaar_legal_validity	right_to_privacy | linking _dbt_to_aadhaar	4%	43%
9	general	transforming_india	4%	47%
10	link_aadhaar_scholarships	aadhaar_enabled_scholarships | exam_participation	4%	50%
11	aadhaar_enrolment	aadhaar_related_documentation	4%	54%
12	aadhaar_public_distribution_system	Food_security | aadhaar	4%	58%
13	politics	political_protest	4%	61%
14	link_aadhaar_voting	voting | voter_id	4%	65%

LDA visualization on with 29 topics:

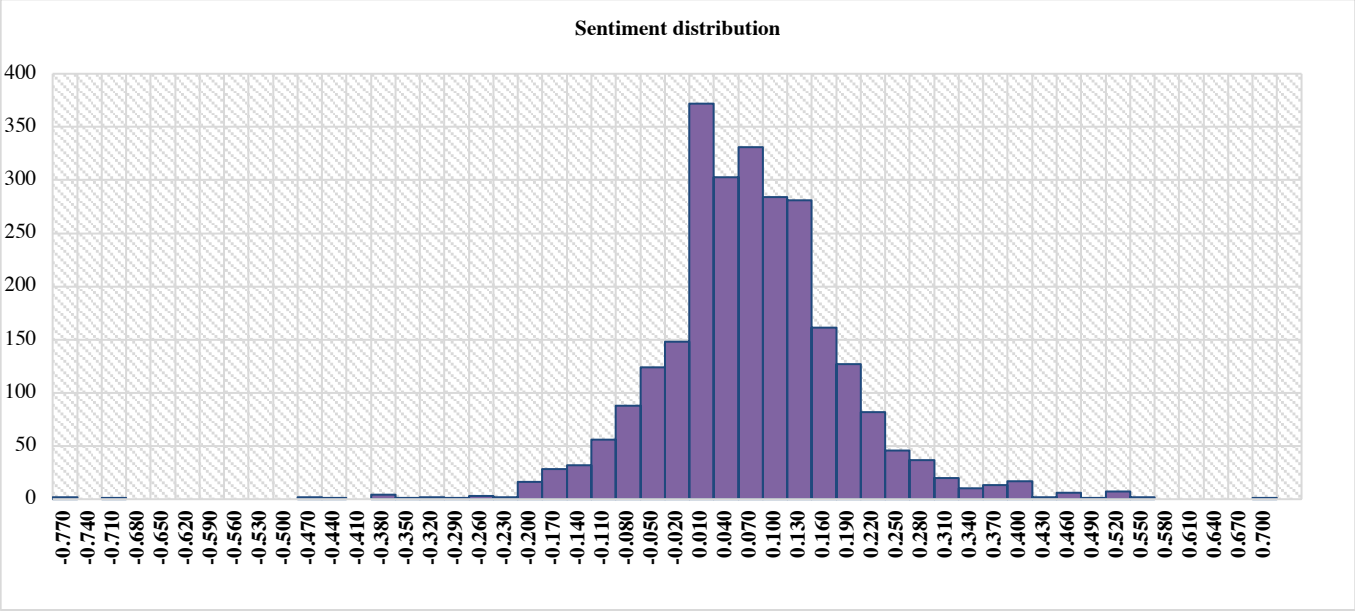
On the visualization tool, we click a circle in the left panel to select a topic, and the bar chart in the right panel displays the 30 most relevant terms for the selected topic. Relevance of a term to a topic is determined given weight parameter, $0 \leq \lambda \leq 1$, as $\lambda \log(p(\text{term} \mid \text{topic})) + (1 - \lambda) \log(p(\text{term} \mid \text{topic})/p(\text{term}))$. The red bars represent the frequency of a term in a given topic, (proportional to p(term | topic)), and the blue bars represent a term's frequency across the entire corpus, (proportional to p(term)). The value of λ is adjusted to rank the terms-- small values of λ (near 0) highlight potentially rare, but exclusive terms for the selected topic, and large values of λ (near 1) highlight frequent, but not necessarily exclusive, terms for the selected topic. For the purpose of interpretation of topics, I have used the words for **$\lambda = 0$, $\lambda = 0.34$ and $\lambda = 1$.**

***(Refer to the HTML attached with the report)**

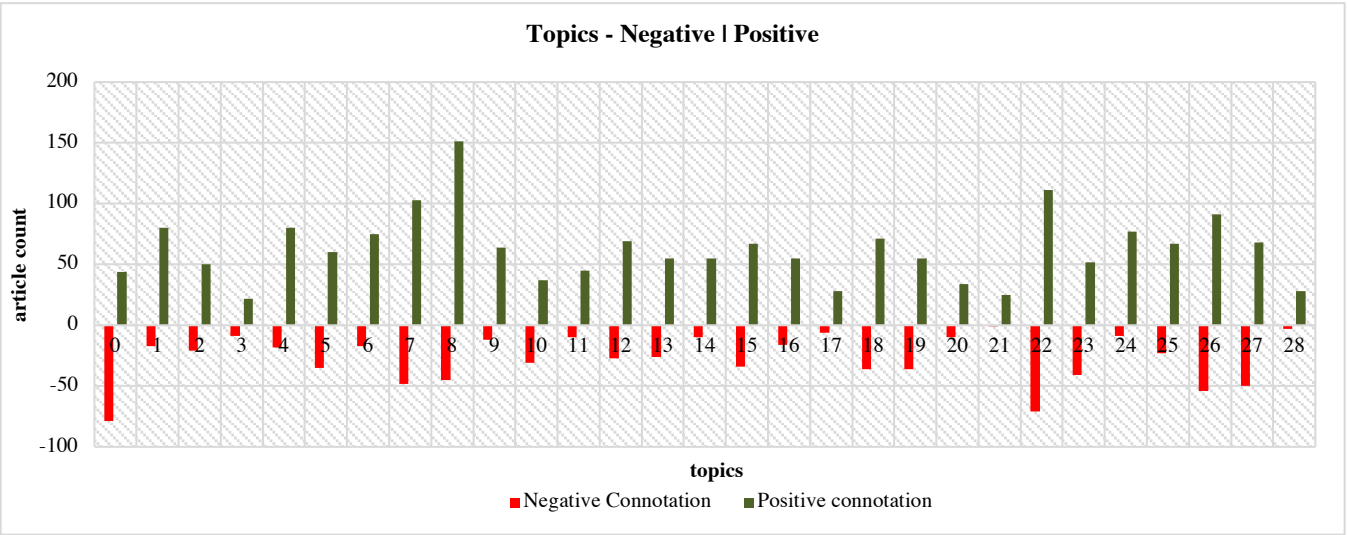


SUPERVISED LEARNING
Sentiment Analysis

Each article consists of a set of positive words , negative words and neutral words. On the basis of these words, we mapped the sentiment polarity of each article in the range [-1,1]. A value less than '0', implies a negative connotation while a value greater than '0' implies a positive connotation. The article's sentiment is mapped to a continuous polarity space. Higher the absolute value of polarity, higher either the negative or positive connotation of the article. Based on the algorithm, both the mean and the median sentiment values are 0.05 to 2 decimal places.



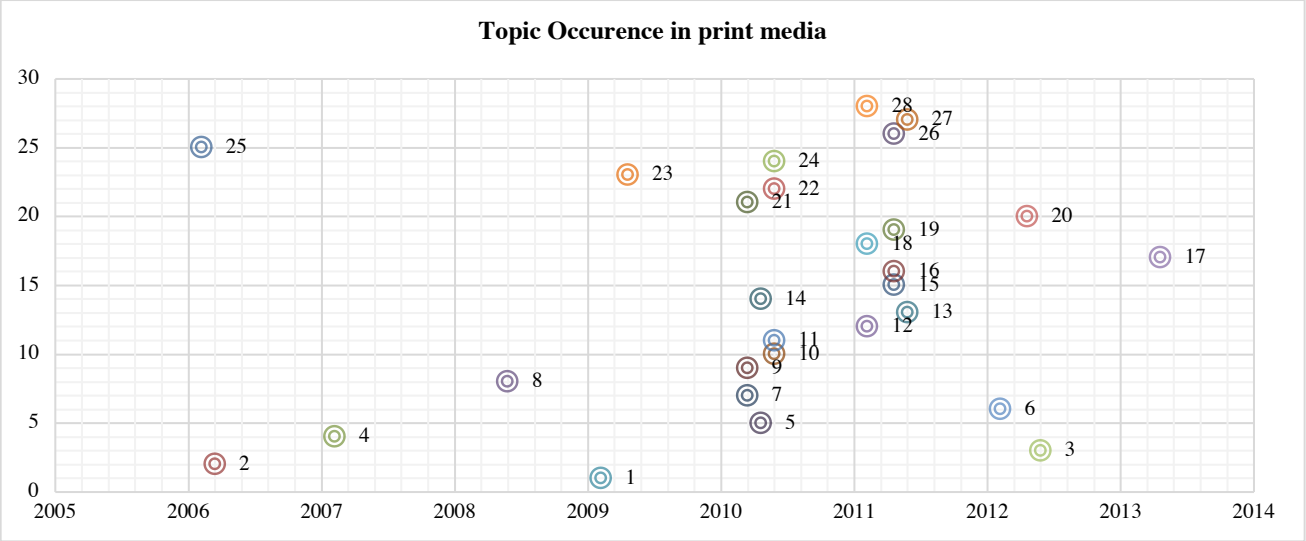
For each topic in the range [0, 28], we plot for each the count of articles with sentiment > than 0 (positive) with green shading , and the count of articles with sentiment < 0 (negative), with red shading.



The following 13 topics are grouped in the decreasing order of percentage of articles with negative sentiments. The table is read as follows: 64% of articles in the Topic 0 category, '**aadhaar_related_crime'** with the subtopic as '**fake_card | misuse** ' display a negative sentiment. These discovered topics revolve around the theme of benefit transfers | forms of subsidy through aadhaar. It can be inferred that the authors of the articles appear to be skeptical about either the success or the implementation of the various schemes under the aadhaar umbrella.

Topic number	Main inferred Topic	Sub topic	% negative
Topic 0	**aadhaar_related_crime**	fake_card | misuse	**64%**
Topic 10	**aadhaar_implementation**	aadhaar_related_operational_concerns	**46%**
Topic 23	**politics**	political_protest	**44%**
Topic 27	**aadhaar_direct_benefit_transfer**	rural_employment | subsidy | scholarship	**42%**
Topic 19	**link_aadhaar_to_other_documents**	aadhaar_made_mandatory_Pan_card_PF_mobile_number	**40%**
Topic 22	**aadhaar_enrolment**	aadhaar_card_application | aadhaar_card_update	**39%**
Topic 26	**aadhaar_direct_benefit_transfer**	LPG _subsidy	**37%**
Topic 5	**aadhaar_public_distribution_system**	food_security	**37%**
Topic 15	**aadhaar_legal_validity**	right_to_privacy | linking _dbt_to_aadhaar	**34%**
Topic 18	**general**	general _infrastructure	**34%**
Topic 13	**aadhaar_direct_benefit_transfer**	crop_loan_waiver | subsidy	**32%**
Topic 7	**aadhaar_enrolment**	aadhaar_authentication_databse	**32%**
Topic 2	**link_aadhaar_bank_accounts**	Financial_Inclusion | Jan_dhan_accounts	**30%**

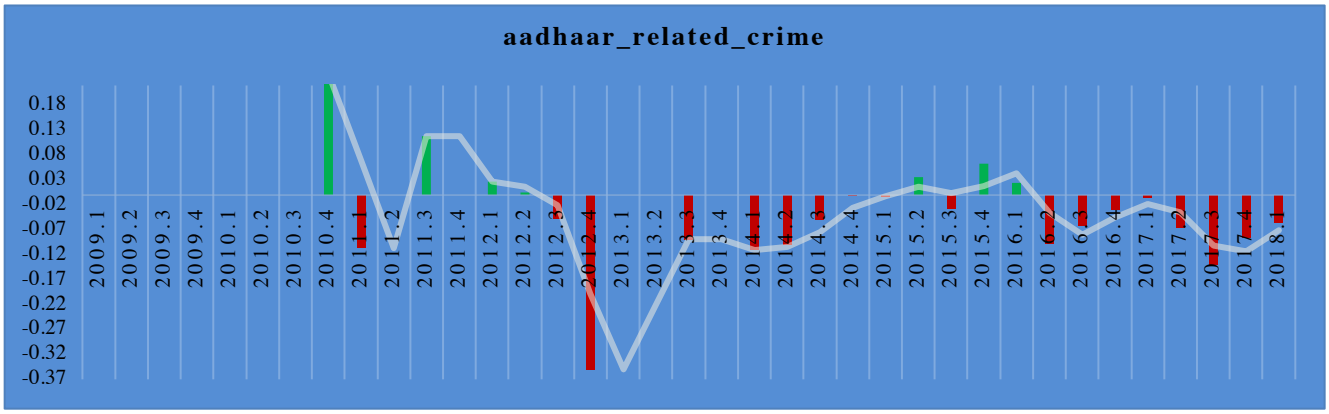
Next, we plot the timeline of emergence of the various discovered topics and observe how certain topics emerge in clusters on the timeline. Topic 8, emerges in the media around the 2nd calendar quarter of 2008. Topic 8 covers articles that describe the onset of the implementation of aadhaar and pilot testing of collecting data for unique identity mapping.



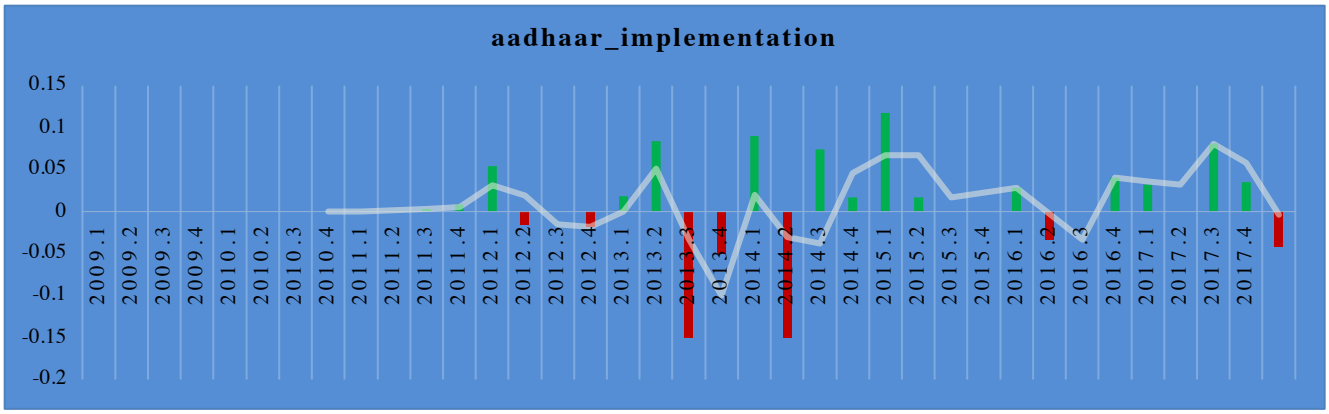
A TIMELINE OF TOPIC WISE ARTICLE SENTIMENT

In this section, we track how the sentiment polarity around a specific topic varies with time. However, this is only the 1st attempt at sentiment polarity. We will further refine the approach to computing sentiment polarity.

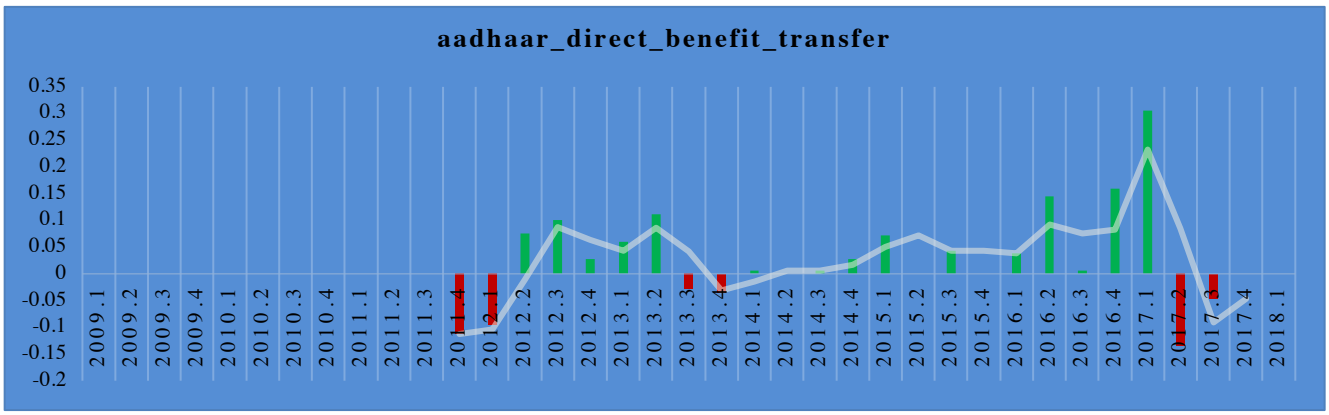
Topic 0 aadhaar_related_crime



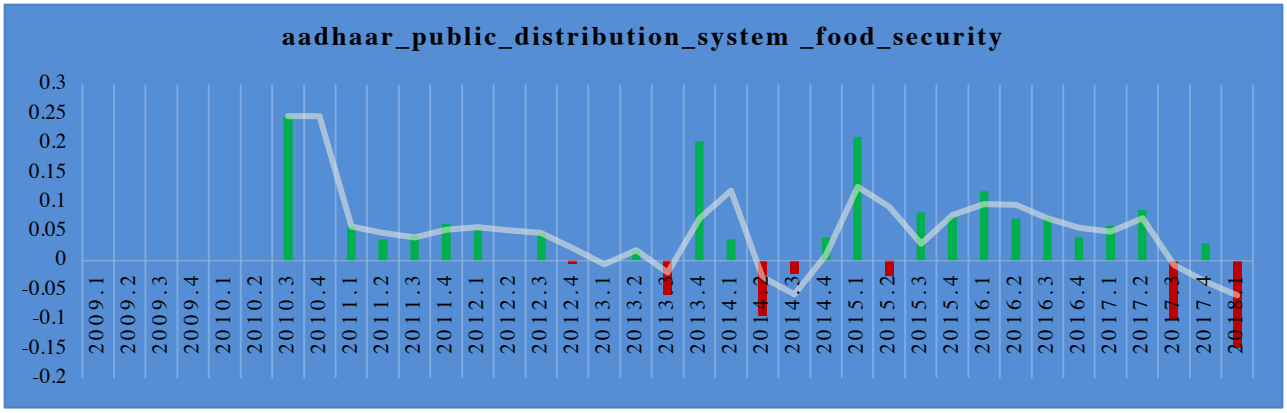
Topic 10 aadhaar_implementation (aadhaar_related_operational_concerns)



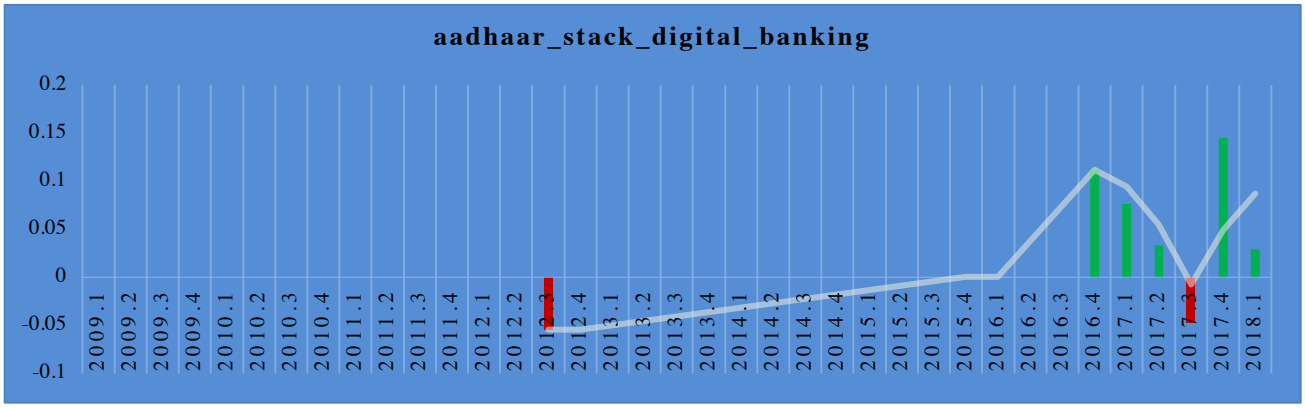
Topic 27 aadhaar_direct_benefit_transfer (rural_employment | subsidy | scholarship)



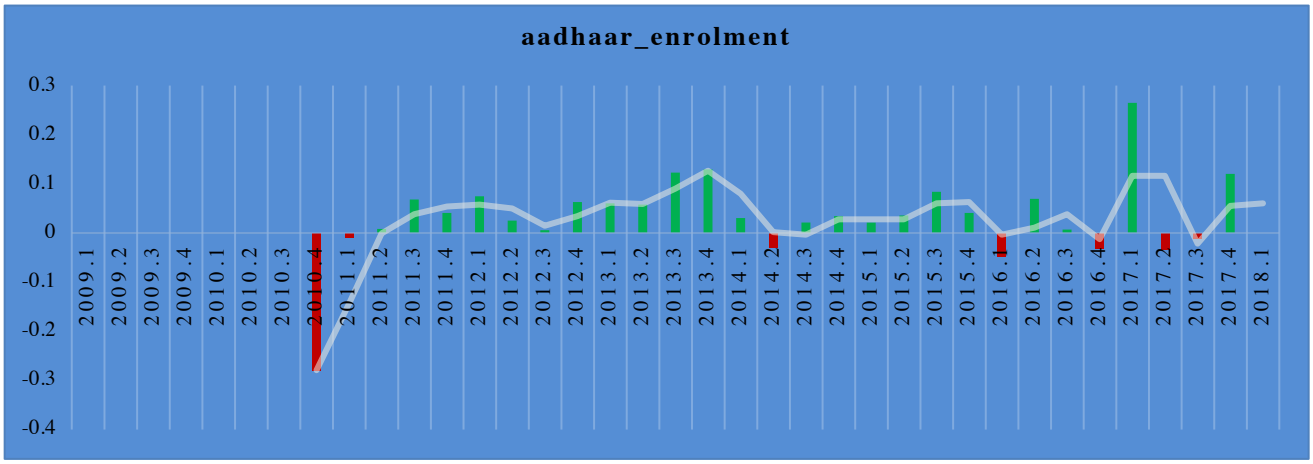
Topic 5 aadhaar_public_distribution_system (food_security)



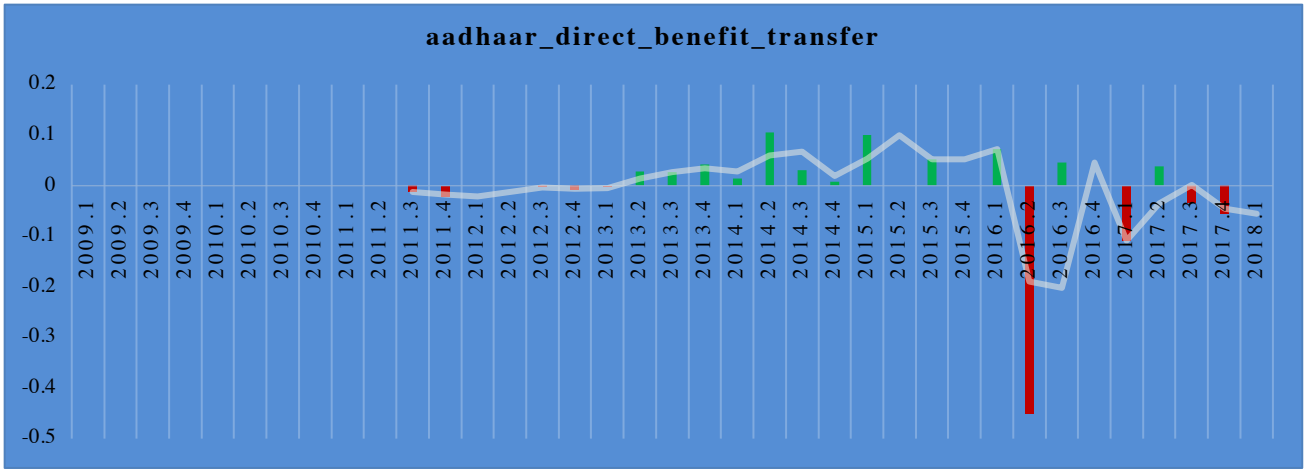
Topic 20 aadhaar_stack_digital_banking (benefits | demonitization | digital_payment | cashless_economy)



Topic 22 aadhaar_enrolment (aadhaar_card_application | aadhaar_card_update)



Topic 26 aadhaar_direct_benefit_transfer (LPG _subsidy)



Continuation

This corpus of this brief are the articles of The Hindu. We intend to extend this work to mining and analyzing media articles from a collection of news papers such as the telegraph for the following purposes; We intend to carefully map the emergence of aadhaar centric mutually exclusive topics over time in different print media together with developing a more robust form of sentiment analysis so as to carefully tag the sentiment polarity around articles.

References

1. L.J.P. van der Maaten. **Accelerating t-SNE using Tree-Based Algorithms**. ***Journal of Machine Learning Research*** 15(Oct):3221-3245, 2014. PDF [Supplemental material]
2. http://scikit-learn.org/stable/
3. https://pypi.python.org/pypi/lda
4. http://textblob.readthedocs.io/en/dev/install.html documentation