PDC Hackathon 2019

•••

Feb 16, 2019

By,
Pramod, Varsha, Prajakta

Petitions

-A formal written request, typically one signed by many people, appealing to authority in respect of a particular cause.

Some of the famous petitions those have changed the world:

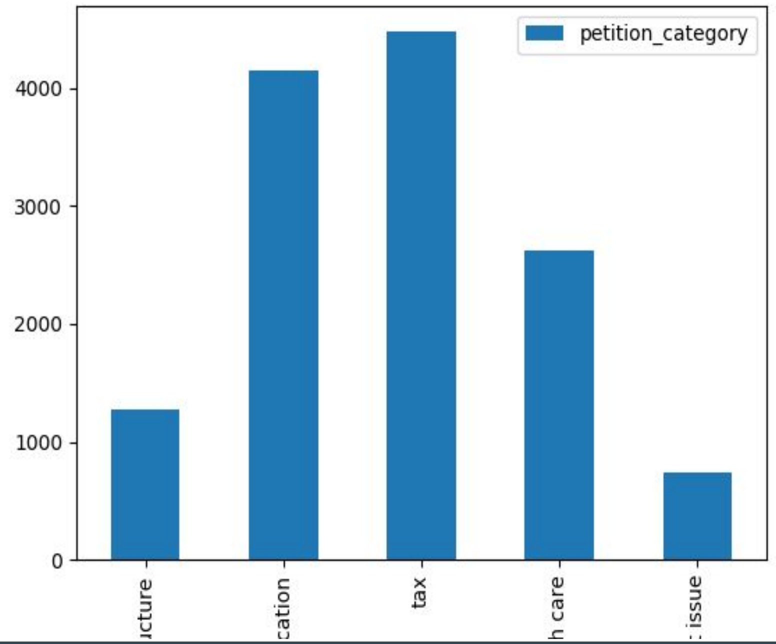
-Give the Meningitis B vaccine to ALL children, not just newborn babies.

"All children are at risk from this terrible infection, yet the Government plan to only vaccinate 2-5 month olds. There needs to be a rollout programme to vaccinate all children, at least up to age 11. Meningococcal infections can be very serious, causing MENINGITIS, SEPTICAEMIA & DEATH."

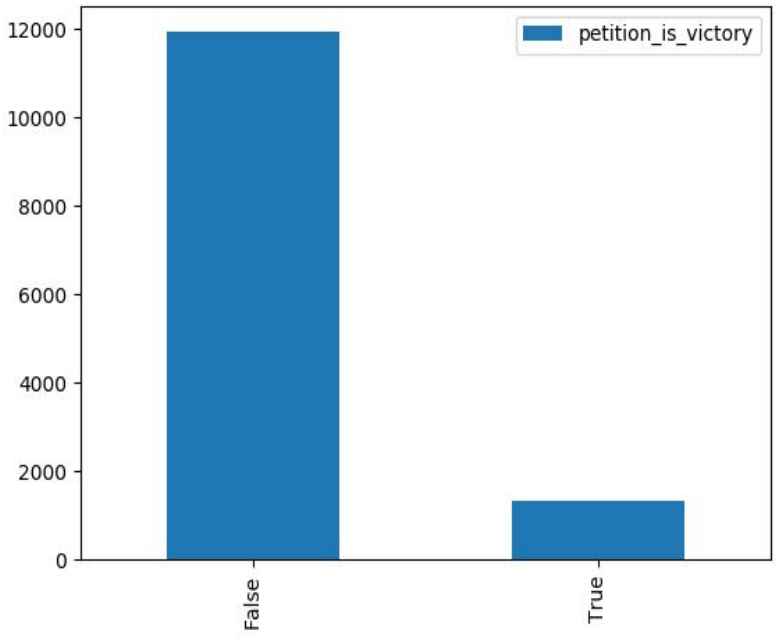
-Vote no on military action in Syria against IS in response to the Paris attacks

Problem Statement

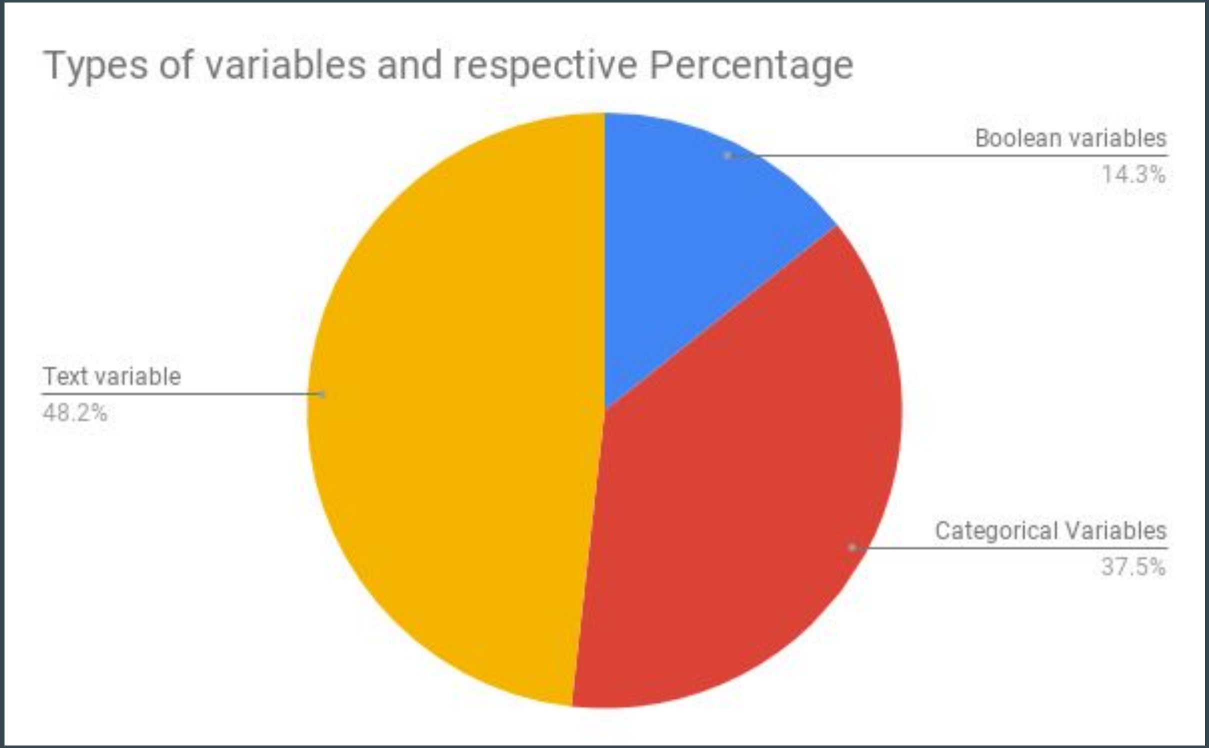
Petition Classification:



Predicting Petition Victory:



Understanding the data



Understanding the data

Boolean Variables(8)

- _source_coachable
- _source_sponsored_campaign
- petition_primary_target_publicly_visible
- _source_discoverable
- petition_discoverable
- petition_primary_target_is_person
- petition_sponsored_campaign
- _source_sponsorship_active

Categorical Variables(21)

- Petition_original_locale
- Petition_petition_status
- Petition_primary_target_display_name
- Petition_primary_target_type
- Petition_primary_target_publicly_visible
- Petition_primary_target_slug
- Petition_primary_target_type
- Petition_user_country_code
- Petition_total_signature_count
- Petition_weekly_signature_cout
- Petition_sponsored_campaign
- petition_user_country_code
- Petition_relevant_location_country_code
- Petition_organization_zipcode
- Petition_organization_state_code
- Petition_organization_state
- Petition_organization_postal_code
- Petition_organization_city
- petition_organization_country_code
- petition_relevant_location_country_code
- petition_petition_status

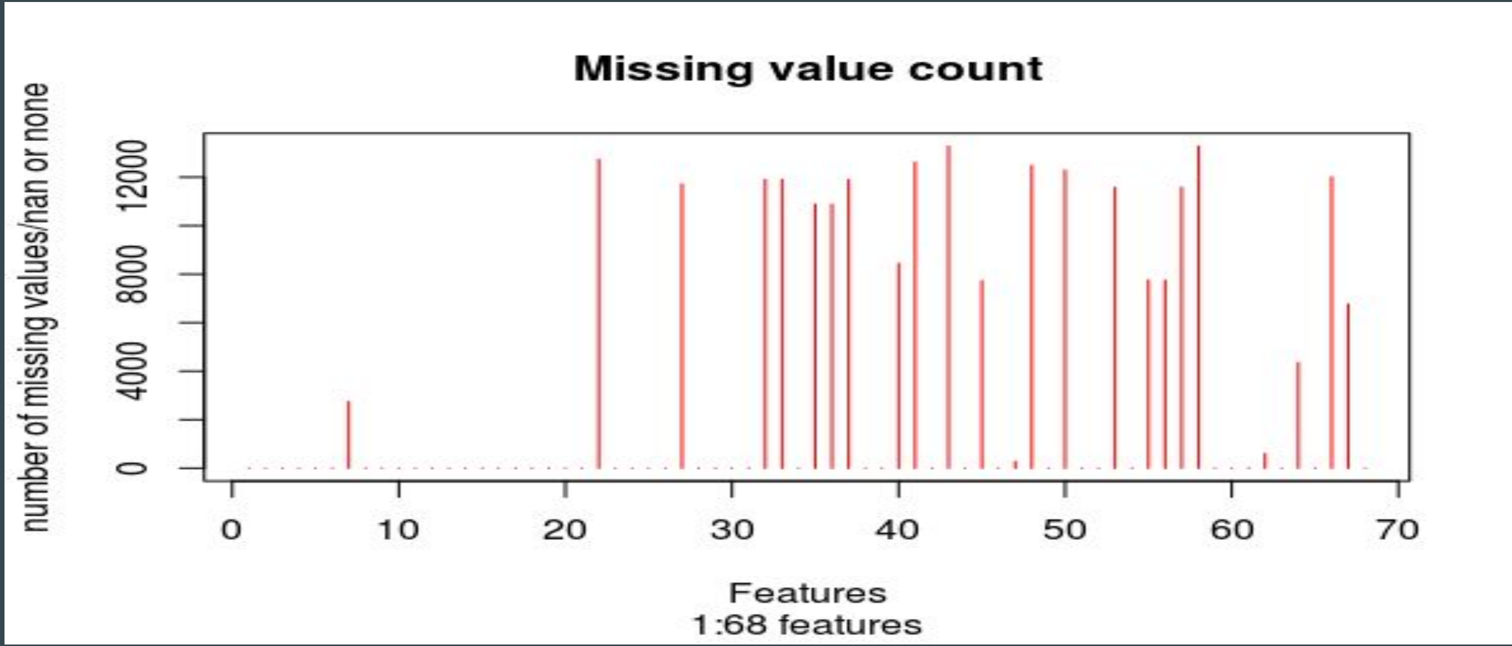
Understanding the data

Text Variables (27)

- _source_ask
- highlight_ask
- highlight_description
- highlight_letter_body
- highlight_targeting_description
- petition_ask
- petition_category
- petition_created_at
- petition_description
- petition_display_title
- petition_languages
- petition_letter_body
- petition_organization_formatted_location_string
- petition_organization_name
- petition_organization_slug
- petition_original_locale
- petition_primary_target_display_name
- petition_primary_target_slug
- petition_primary_target_type
- petition_published_at
- petition_slug
- petition_targeting_description
- petition_title

Feature Engineering(for both problems)

- **Based on NA count proportion with training data size:**



Feature Engineering:

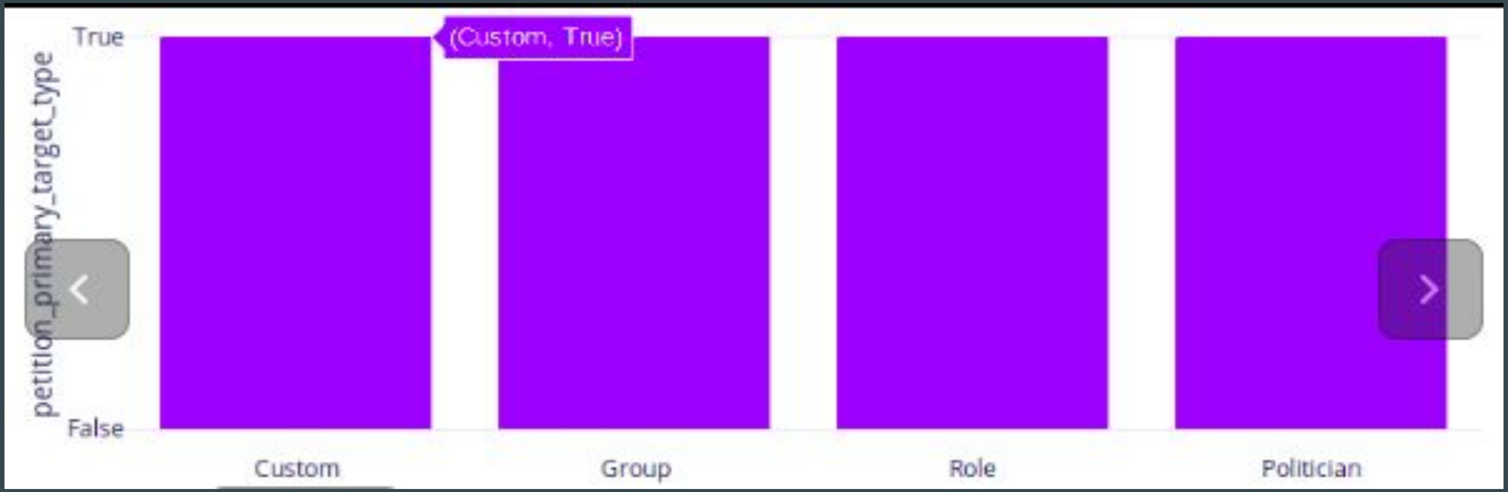
petition_organization_zipcode	13276
petition_organization_state_code	13276
petition_organization_state	12733
petition_organization_postal_code	12627
petition_organization_city	12476
petition_goal	12292
petition_organization_non_profit	12006
petition_primary_target_additional_data_title	11901
petition_primary_target_description	11897
petition_primary_target_email	11897
petition_primary_target_locale	11722
petition_primary_target_summary	11582
petition_primary_target_verified_at	11579
petition_relevant_location_city	10886
petition_relevant_location_lat	10886
petition_relevant_location_lng	8439
petition_relevant_location_state_code	7755
petition_restricted_location	7755
petition_user_description	7726
petition_user_state_code	6764

These features excluded due to having greater than 50% missing values

Feature Engineering

Removal of few features based on Based on data distribution plots:

-For each unique value has came in both class for around same number of emails. Has no significance with respect to is_victry.



Petition Category Classification

-The petitions are categorized into below 5 categories:

Tax	**4475**
Education	**4151**
Health Care	**2625**
Infrastructure	**1279**
Environment Issue	**746**

Initial Approach-Petition Classification

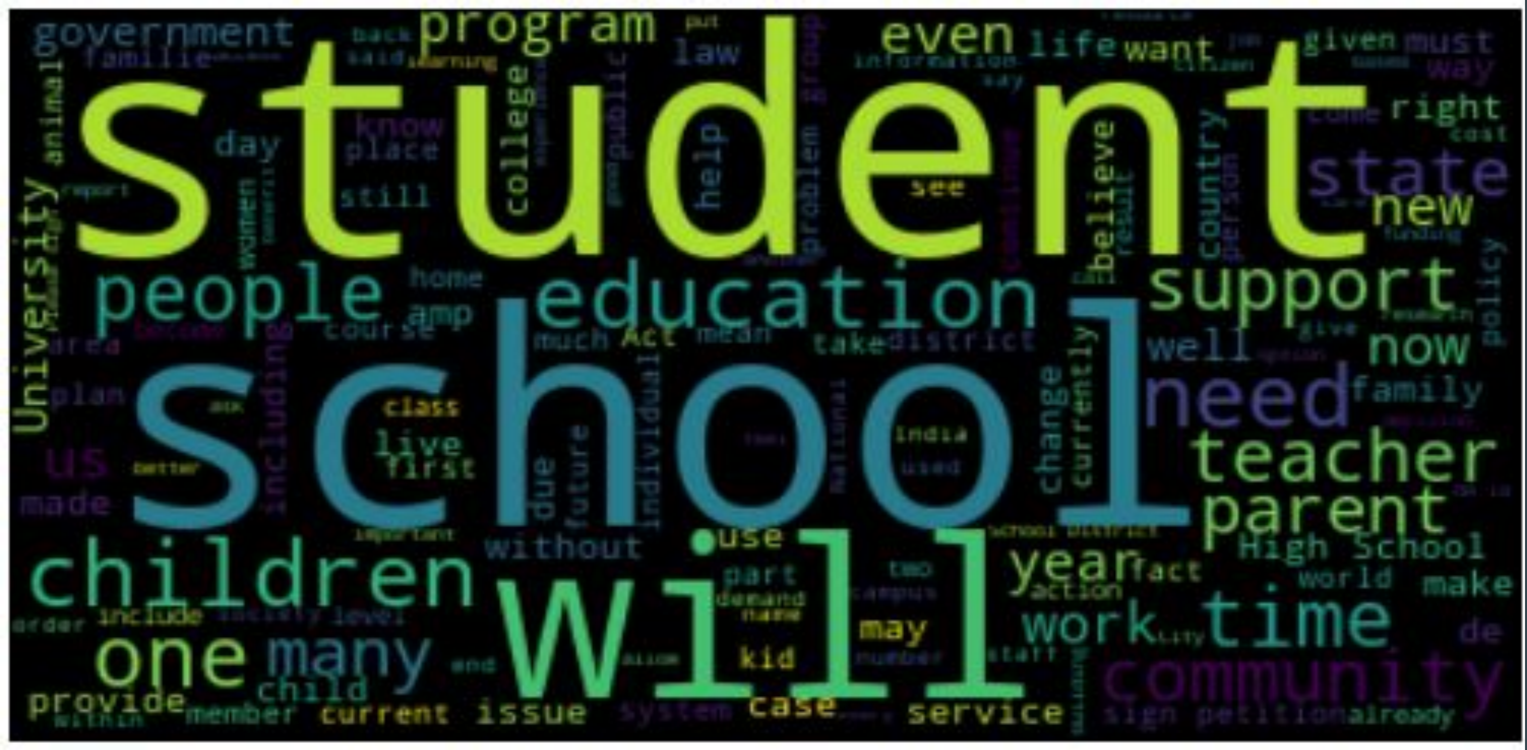
-After understanding the problem, it looks like this is text classification problem.
-Below features will be useful for categorizing the petitions:

- -petition_letter_body
- -petition_description
- -petition_display_title
- -petition_title
- -petition_targeting_description
- -highlight_ask
- -highlight_description
- -highlight_letter_body
- -highlight_targeting_description

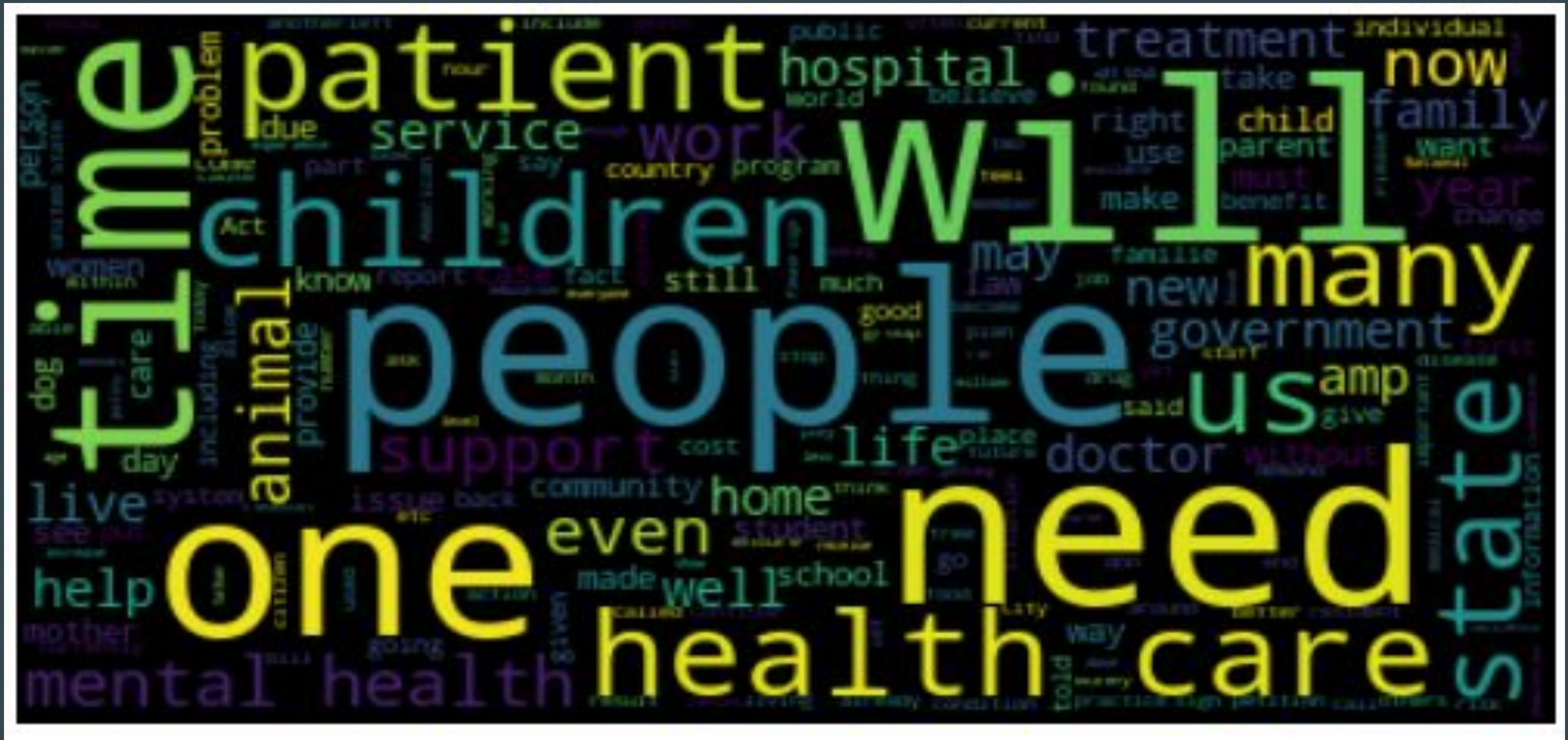
-Feature Generation Techniques:

- -Bag of Words
- -Tf-idf
- -Word2Vec(wikipedia trained, trained on petition data)
- -Sentence Embedding

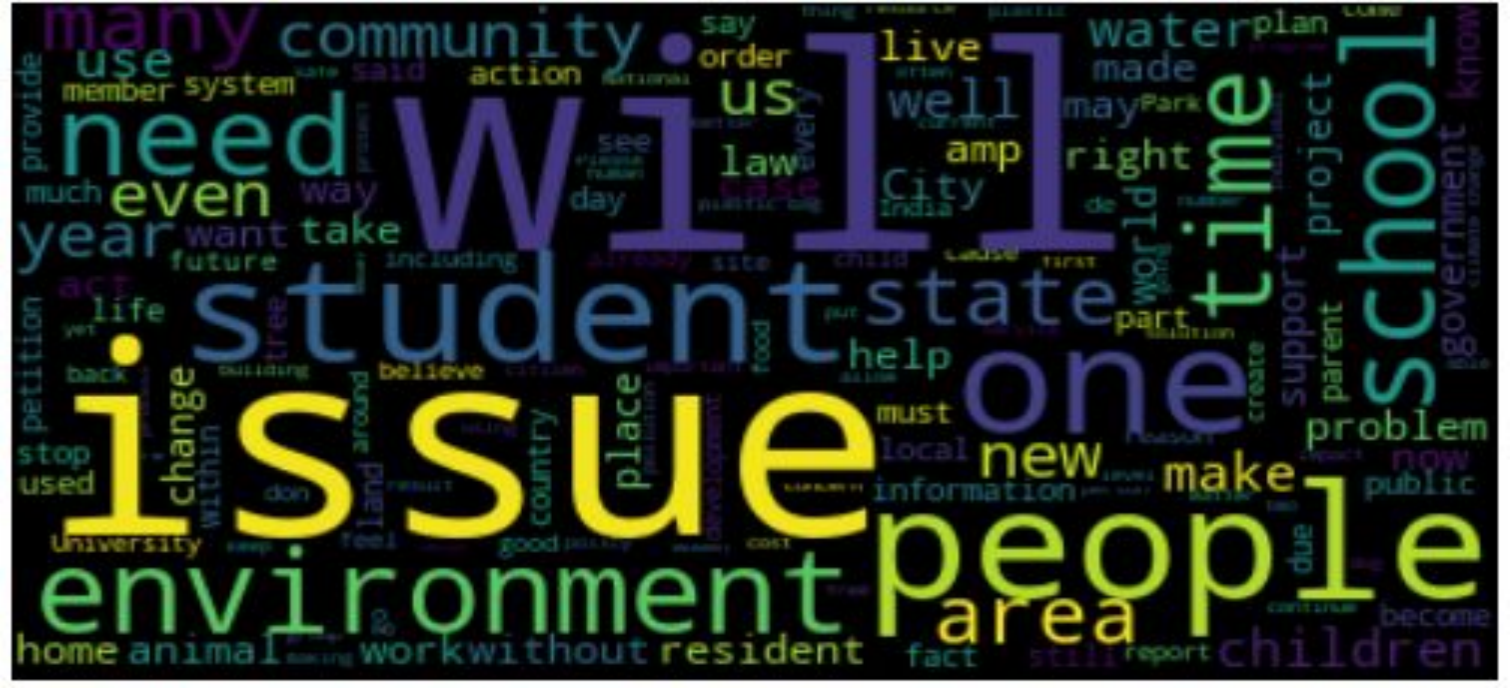
WordCloud For Education



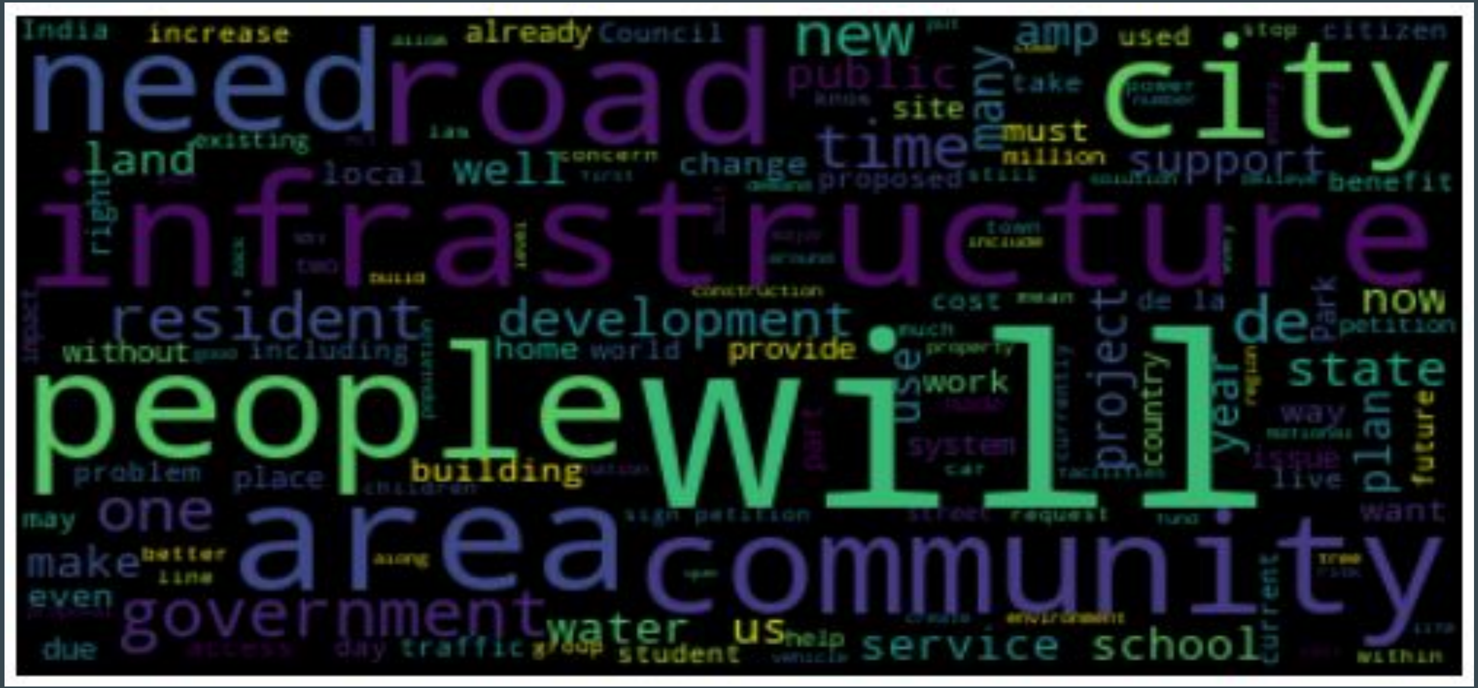
WordCloud For Health Care Petitions



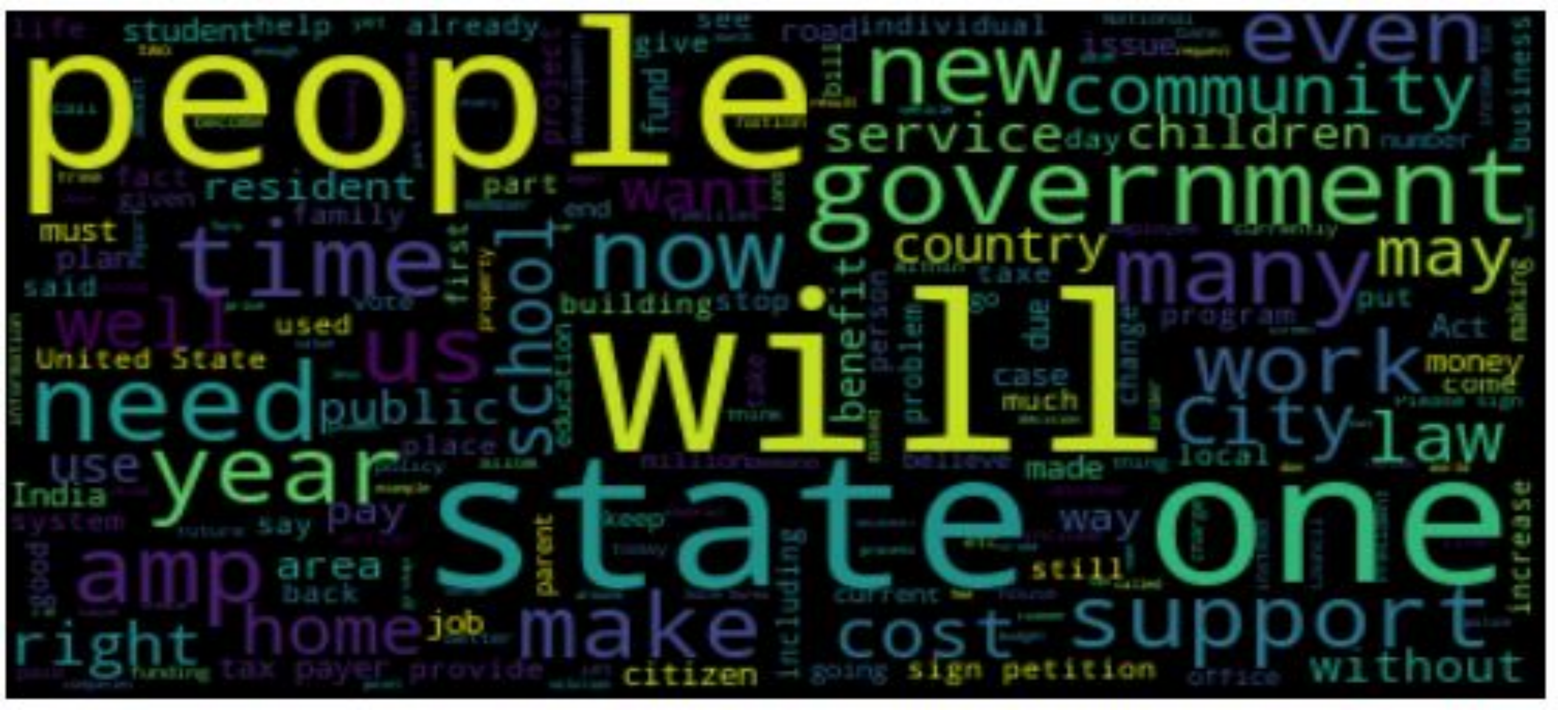
WordCloud For Environment Issue



WordCloud For Infrastructure



WordCloud for Tax Petitions



Initial Results-Petition Classification

Below are the best results achieved so far

Technique:

1. CountVect/Tf-idf
2. Feature tried: petition_description, petition_letter_body, petition_targeting_description
3. Stratified split applied:
   a. Train_data : 75%
   b. Validation_data: 25%
4. Best results achieved using **petition_descrption,**
   a. CountVect(n_gram=(1,2), max_features=3000, stop_words='english', min_df=5,)
   b. RandomForest with default parameters
   c. F1 score on test data: 0.7983
   d. test accuracy: 0.8849

Confusion Matrix

	Education	Environment issue	Health care	Infrastructure	Tax
Edu	946	3	46	9	34
Env	51	109	11	6	9
Health C	81	5	545	6	19
Infra	37	4	7	247	25
Tax	56	4	28	11	1020

Petition Classification: Final Approach

- We tried CountVectorizer on each of text column mentioned in slide number 9
- For highlight_ask , we got around 99% F1 score
- So we checked the text manually. Below are some of the samples:
    - "['Use public <mark>infrastructure</mark> for state sponsored conferences']"
    -"['Pressure State Legislature for more Police Department training and <mark>education</mark>']"
    -"['State of Virginia: Institute a wealth <mark>tax</mark> on the rich. Inequality is MORALLY WRONG.']"
- We saw that highlight_ask of each of the petition. We found that Each of the target information is marked in this column
- We applied regular expressions and extracted the marked text.
- We then designed rule based classifier for classifying the petition type.

Results on 25% Stratified Validation Data:

- Accuracy: 100%

- F1 Score: 100%

Advanced NLP Techniques

-In Case, we don't want to use rule based classifier.
-The same can be used by Bag of words with the use of limited vocabulary.
-In case we want to go for state of art NLP techniques, below techniques can be used:

- Average of Word2vec/Glove
- Weighted Average of Word2vec(check the paper)
- Word2Vec can be trained on petition data as well, as there is enough text available.
- Sentence Embedding by Universal Sentence Encoder.
- Transfer Learning using ULMFit
- Language Models for vectorization. Below are the some of state of art LMs:
    - ELMo(Deep contextualized word representations)
    - BERT(Pre-training of Deep Bidirectional Transformers for Language Understanding)
    - Pre-trained models are available of language models
    - Or we can fine-tune these language models

Petition_is_Victory: Initial Approach

-The data is imbalanced for this problem

-Below is the distribution:

False	11938
True	1338

-There are around 10% petitions are actually won.

-Still 10% of instances are enough for modelling. The problem becomes difficult if the ratio becomes 1% vs 99%. So with the current amount of data, we think the problem is solvable by using techniques like undersampling, upsampling(SMOTE), bagging, ensemble.

Excluded due to perticular reason as stated(Vectory)

- Filename: Doesn't give any information related to problem statement
- Petition_user_city: Not giving any relevant info(3799 unique values and nan)
- _source_sponsorship_active: only 16 out of 13227 are false 2300:None
- Petition_sponsored_campaign: only 19 false
- _source_sponsorship_campaign: only 19 false
- Petition_discoverable: All True
- _source_discoverable: ALL True
- petition _id: No use in modeling
- Petition_petition_status: this seems to be very important but it can lead model to behave improperly on test data
- For deciding victory of petition as per the domain knowledge we have excluded these columns as these all are text data, mainly scores, number of signatures are important:
  - Petition_title,petition_primary_target_slug, petition_targeting_description,
  - petition_organization_name, petition_user_country_code,
  - petition_organization_formatted_location_string, _source_country_code,
  - Petition_relevant_location_country_code, petition_primary_target_display_namehighlight_ask, highlight_description, highlight_letter_body, highlight_targeting_description, petition_ask, petition_category,petition_created_at,petition_description,petition_display_title, petition_languages Petition_letter_body,petition_organization_country_code,petition_organization_id,petition_organization_slug,petition_original_locale,petition_petition_status,petition_published_at,petition_slug,_source_user_state,_source_ask

Approaches to handle data imbalance

- Undersampling majority class examples
- Bagging with undersampling(Multiple models can be trained on random sample of majority class vs minority class, and majority vote can be used.)
- Oversampling minority class examples(SMOTE)
- Giving weightage(Penalizing the majority class based on distribution)

Due to time restrictions only first approach we tried

Models used, results(with default parameters)

XGBoost

-F1_Score: 0.7445

-Accuracy: 0.71041

-Confusion matrix:

Actual\Predicted	False	True
False	255	191
True	65	373

Random Forest

-F1_Score: 0.7409

-Accuracy: 0.6946

-Confusion matrix:

Actual/Predicted	False	True
False	228	218
True	52	386

Logistic Regression

-F1_Score: 0.6801

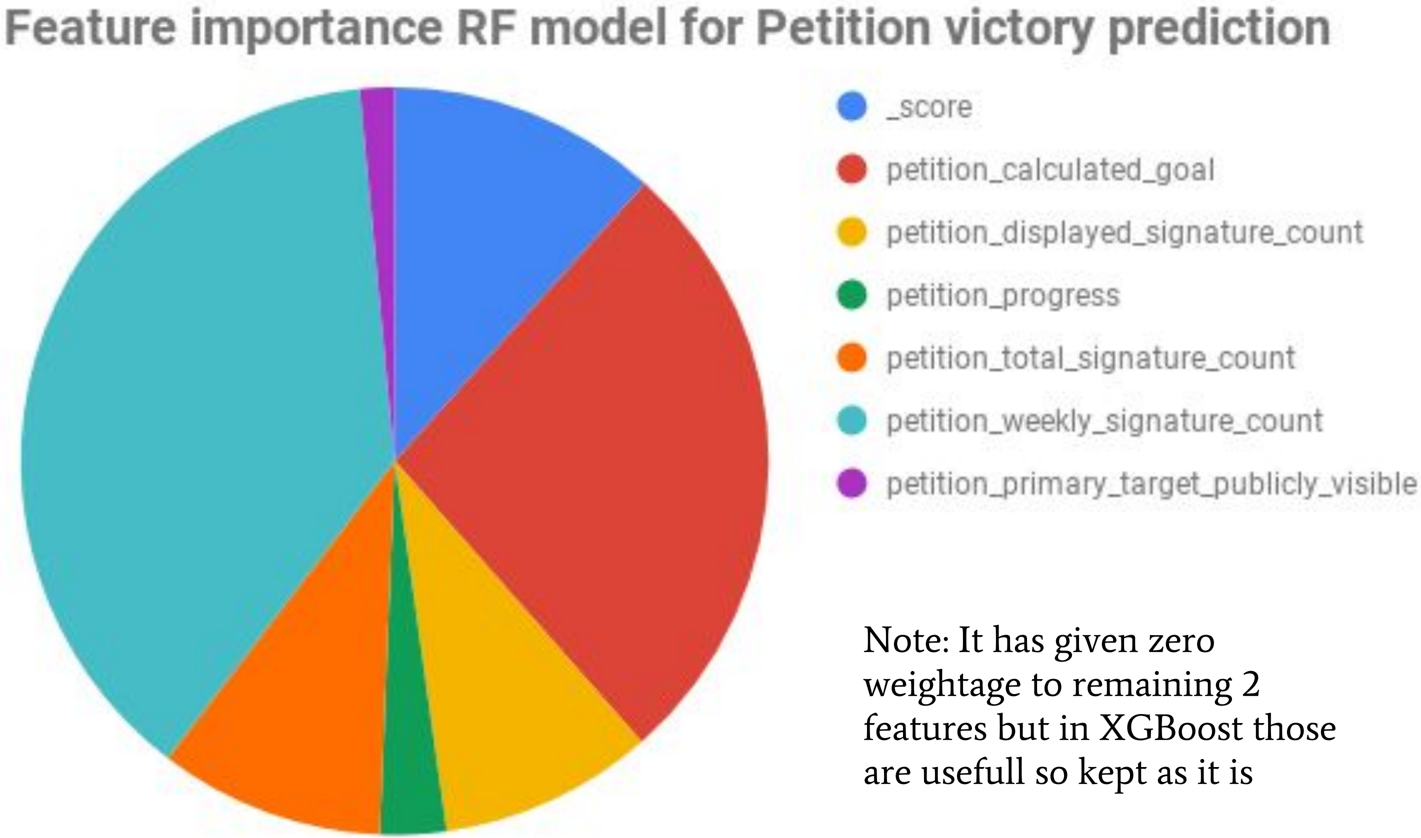
-Accuracy: 0.5339

-Confusion matrix:

Actual/Predicted	False	True
False	34	412
True	0	438

Features Used:

1. _score
2. Petition_calculated_goal
3. Petition_displayed_signature_count
4. Petition_primary_target_publicly_visible
5. Petition_primary_target_type
6. Petition_progress
7. Petition_total_signature_count
8. Petition_weekly_signature_count
9. _source_coachable



Note: It has given zero weightage to remaining 2 features but in XGBoost those are usefull so kept as it is