AUTOMATIC EXTRACTION OF COORDINATES FROM ATELS

RND PRESENTATION

Naman Jaswani

namanjaswani@iitb.ac.in

193070052

WHAT IS AN ATEL ?

- The Astronomers Telegram (ATEL) is an internet based short notice publication service for quickly disseminating information on new astronomical observations.

Telegram Index

Telegrams Posted Within the Last 30 Days (All)

70 Selected of 13836 Telegrams

13836	Small apparent outburst of comet 115P/Maury	Michael S. P. Kelley, Quanzhi Ye, Dennis Bodewits on behalf of the Zwicky Transient Facility Collaboration;... *26 Jun 2020; 20:24 UT*
13835	New giant radio flare from Cyg X-3, correlated with X-rays and gamma-ray flares	S. A. Trushkin, N. N. Bursov, A. V. Shevchenko, N. A. Nizhelskij, P. G. Tsybulev, A. N. Borisov *26 Jun 2020; 20:19 UT*
13834	The New Variable Nebula in Cepheus - A Sign of a Recent FUOr Event	Bringfried Stecklum *25 Jun 2020; 23:36 UT*
13833	H-alpha confirmation and additional optical photometry of the candidate nova M31N 2020-06a	A. Valcheva, M. Minev, E. Ovcharov, P. Nedialkov *25 Jun 2020; 20:40 UT*
13832	New Variable Nebula in Cepheus	G. Borisov, D. Denisenko *25 Jun 2020; 18:16 UT*

Small apparent outburst of comet 115P/Maury

ATel #13836; *Michael S. P. Kelley (U. Maryland), Quanzhi Ye (U. Maryland), Dennis Bodewits (Auburn U.) on behalf of the Zwicky Transient Facility Collaboration; Brian Skiff (Lowell Obs.)*
*on **26 Jun 2020; 20:24 UT***
Distributed as an Instant Email Notice Comets
Credential Certification: Quanzhi Ye (qye@umd.edu)

Subjects: Optical, Comet

Tweet

We report the discovery of an apparent outburst of comet 115P/Maury with the Zwicky Transient Facility (ZTF; Bellm et al. 2019, PASP, 131, a8002). Based on automated processing of nightly ZTF data, a small outburst occurred between 2020 Jun 19.36 and 2020 Jun 23.34 UTC. The event is initially apparent in two exposures in both photometry and morphology as a compact source centered on the coma. We followed-up on the event with images from the Lowell Observatory 0.8-m telescope taken 2020 Jun 24.20 UTC through an R-band filter.

Photometry measured in a 5" radius aperture (4100–6500 km) is tabulated below, calibrated to the PS1 catalog (ZTF; Tonry et al. 2012, ApJ 750, 99), or ATLAS-RefCat2 (Lowell; Tonry et al. 2018, ApJ 867, 105). We estimate the $g–r$ color to be 0.49±0.01 mag, and fit a model to the effective r-band lightcurve: m = 11.4 + 5 log10(Delta) + 14 log10(rh) + Phi(phase), where Phi is the Halley-Marcus phase function (Schleicher & Bair 2011, AJ 141, 177) expressed in magnitudes (RMS of 0.06 mag). Based on this fit, the outburst was at least –0.3 mag. No other events are apparent in our data.

The comet is at low Galactic latitudes, but the ZTF image differencing pipeline mitigates stellar contamination, and the photometric apertures on Jun 23–26 avoided any significant stars. Monitoring of the comet is encouraged in case this event is followed by a larger one in the near future.

PROBLEMS FACED BY ASTRONOMERS

- In observational astronomy, one cannot wait for a proper peer review process before publishing a discovery. This is because many times the observations made are short in time scale, and by the time peer review is done, the event has already faded away.
- There has been significant number of cases when astronomers report the same discovery without reading the circulars.
- Also many times, astronomers might want to access all circulars which referred to a particular location in space.

OBJECTIVE

- *The project aims at extracting location of observed objects from ATels in an automated way. This will further be used to create an interface to query all circulars which referred to a particular location.*

OUR FIRST APPROACH: THINKING IT IN DEEP LEARNING WAY !

- Our first approach was to look at this problem as a NER task
- Named Entity Recognition(NER) aims at recognising various *tokens as entities. These could be* ***PERSON****,* ***DATE****,* ***LOCATION****,* ***ORGANISATION****, etc.*
- *In our case it would consist of only two entities namely* ***COORDINATE****, and* ***O*** *(not a coordinate).*

NER MODELS

- After labelling dataset in LightTag, I generated dataset in NER required format.
- Passed the dataset from BiLSTM model but it didn’t train ! (considering the small dataset we had which would be used to train embedding layers parameters as well as model parameters)
- Then we thought of using BERT language model. But we could not find any pretrained vocabulary for our context (scientific text).
- Even if we try to train our own vocabulary from scratch, we didn’t had any suitable tokenization as our text had various symbols (!,@,#,$,%,&) which most of the tokenizations would split on.

FORMATS

Category 1: Object Name Format

Ehhmmss.ss+ddmmss.s	J180420.99-293108.9

Category 2: Remaining Formats

RA (hours)	Dec (deg)	Example
hh:mm:ss.ss	dd:mm:ss.ss	18:04:20.99 -29:31:08.9
hh mm ss.ss	dd mm ss.ss	18 04 20.99 -29 31 08.9
hh:mm:ss.ss,	dd:mm:ss.ss	18:04:20.99,-29:31:08.9
00h 00m 00.0s	00d 00m 00.0s	18h 04m 20.99s -29d 31m 08.9s
00h00m00.0s	00d00m00.0s	18h04m20.99s -29d31m08.9s
00h00m00.0s	00°00'00.0"	18h04m20.99s -29°31'08.9"
hh.hhhhhhh	dd.ddddddd	18.072497 -29.519139

SO WHY NOT ONLY USE REGEX (REGULAR EXPRESSION)

- False positive with date-time : ‘INTEGRAL performed a target-of-opportunity observation of the colliding wind binary eta Carinae from 2019-12-06 13:10:19 to 2019-12-07 09:28:36 (UTC)’
- Extracting coordinates in decimal format : ‘Object TXS 0358+210 with radio coordinates (J2000) R.A.: 60.43819 deg, Dec.: 21.17461 deg (Beasley et al. 2002, ApJS 141, 13’
- False positive in Object name format: ‘We obtained optical spectra covering the range 5000-9400 A with the 3.58m TNG telescope equipped with LRS at Observatorio del Roque de los Muchachos’

MODIFICATION TO REGEX APPROACH

- Classifier A : Classify a sentence containing object name format match as a false positive or true positive. *Example* : ‘We would like to thank the Swift team and other colleagues for providing continuous Swift monitoring observations of MAXI J1820+070 to the public.’
- Classifier B : Classify a sentence as having coordinate information (in a format other than the object name format). *Example* : ‘Object TXS 0358+210 with radio coordinates (J2000) R.A.: 60.43819 deg, Dec.: 21.17461 deg (Beasley et al. 2002, ApJS 141, 13).’

PROBLEM WITH CLASSIFIER-A

- Dataset generation : Used regex and iterated through CSV of ATels to interactively label each match as TP or FP.
- Huge class imbalance with 484 positive examples while only 50 negative examples
- Without any classifier the accuracy was 92 percent with zeros false negatives and 8 percent false positives.
- So we discarded classifier A.

DATASET FOR CLASSIFIER-B

- On close look, all formats except object were used in same content. This is because object name format is sometimes used to even refer to an object as its name while the other formats are just its location.
- While creating dataset for classifier-B we got huge false positives for decimal format only. So we decided to *use data created for other formats except decimal format to train the model, and then do inference on all formats including decimal format.*
- i.e. Train data : Total data - (Object name format + Decimal format)

CLASS IMBALANCE IN DATASET FOR CLASSIFIER-B

- 286 positives and 117 negatives
- Explicitly added more negative examples from decimal format.
- Decimal format chosen for new negative examples(decimal pair regex) because we don’t want our model to give false positive on numbers other than coordinates in decimal format.

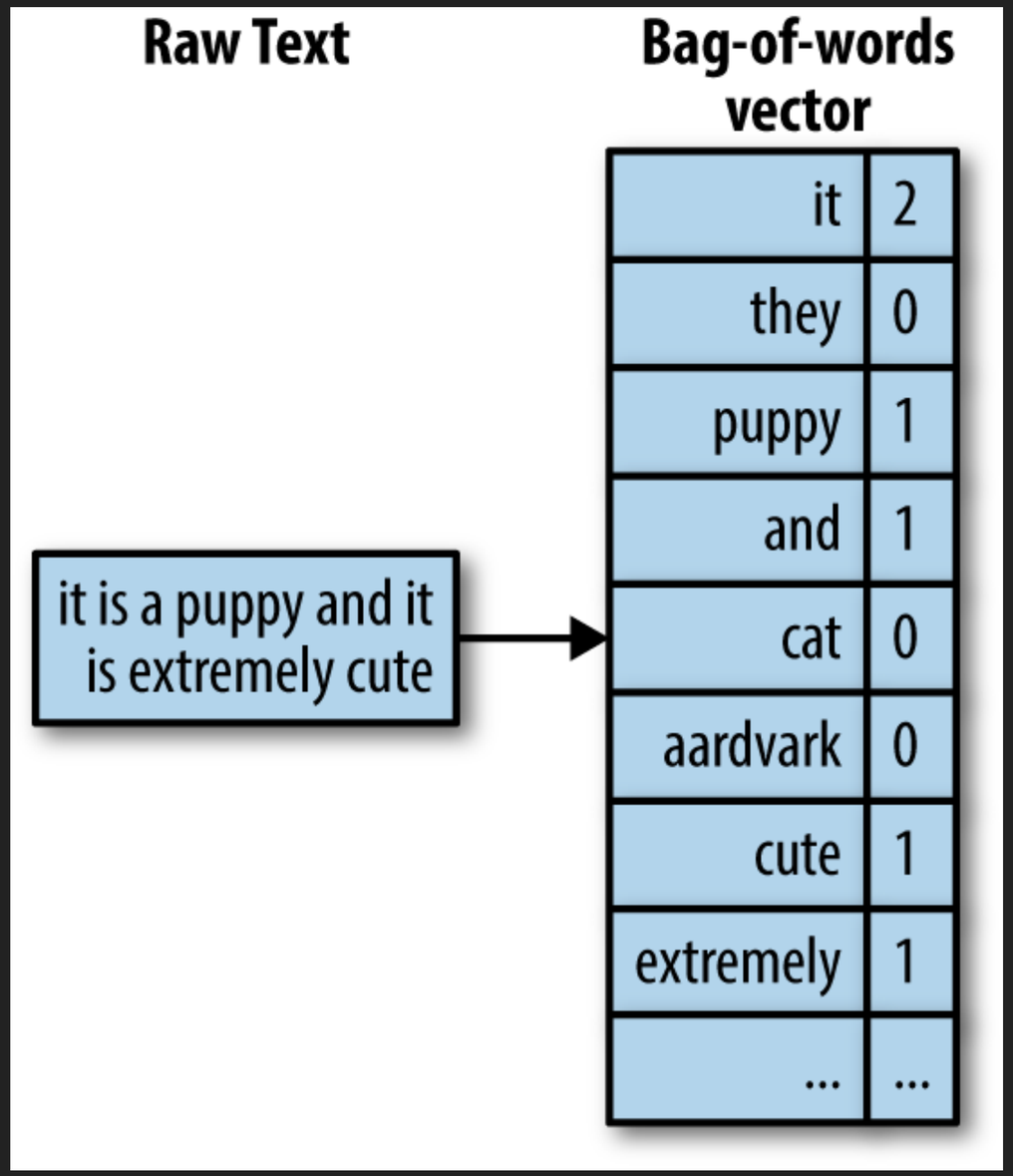
PREPROCESSING

- All characters except alphabets and ‘.’, were removed.
- All characters were converted to lower case and all stop words were removed.
- There are more details of the pre-processing specific to our data

ML PIPELINE

- Bag of words representation using Countvectorizer
- TF-IDF transform (acts as a weight).
  Formula = TermFrequency*InverseDocumentFrequency
  TermFrequency(word i) = # i in sentence / # words in sentence
  InverseDocumentFrequency = log(# sentences/ # sentences containing i)
- Linear SVM classifier

Bag of words representation based on term frequency(tf) or count approach



TF-IDF approach to reflect how important a word is to a document in a collection or corpus

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

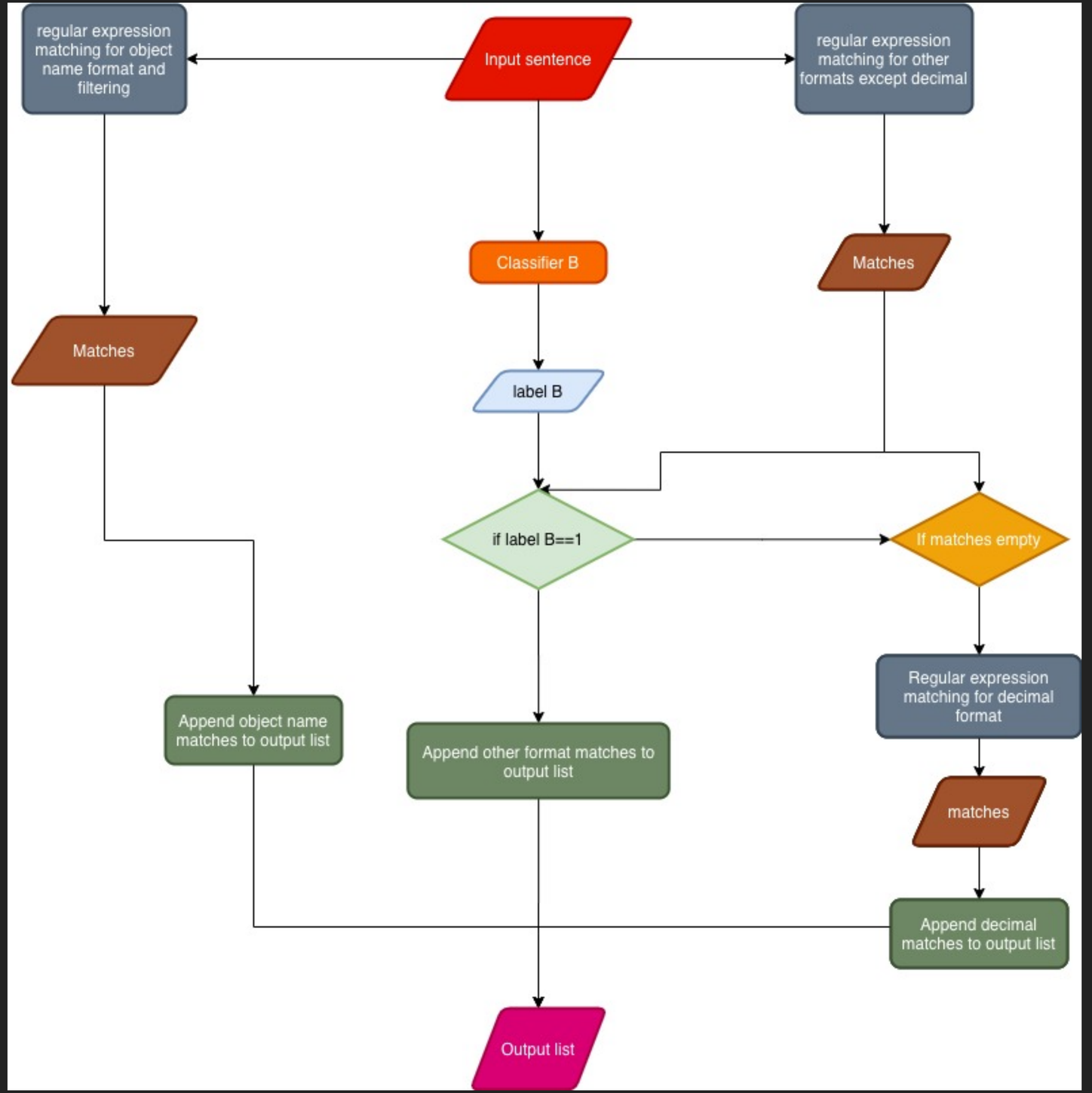
Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

Pipeline


regular expression matching for object name format and filtering
Input sentence
regular expression matching for other formats except decimal
Classifier B
Matches
Matches
label B
if label B==1
If matches empty
Append object name matches to output list
Append other format matches to output list
Regular expression matching for decimal format
matches
Append decimal matches to output list
Output list

Results

Test Results

Accuracy: 95.18%

	Precision	Recall	F1 Score
0	0.96	0.95	0.96
1	0.94	0.95	0.95

Confusion Matrix

TP: 138	FP:7
FN:6	TN:119

Train Results

Accuracy: 99.23%

	Precision	Recall	F1 Score
0	1	0.99	0.99
1	0.98	0.99	0.99

Confusion Matrix

TP: 240	FP:2
FN:1	TN:162

RESULTS

- Model is able to catch false positives:
  (Although the following Sentences gave match with regex, model labelled them as 0)
  1. We obtained 8.6 ksec of NICER data between 2018 November 14 at 22:11:16 UTC and 2018 November 15 at 15:36:13UTC

  2. The object was observed at 17.0-17.2m during 9 minutes from 22:15:26 to 22:24:30 UT on 2018-04-26

- Model is able to extract decimal formats:
  (Although the model was trained on formats other than decimal, it is able to identify sentences which contain coordinates in decimal format.)

  1.Based on the pointing direction of Astrosat at the time of the GW event (RA = 189.2, DEC = 62.3), the FRB was 157 degrees off axis.

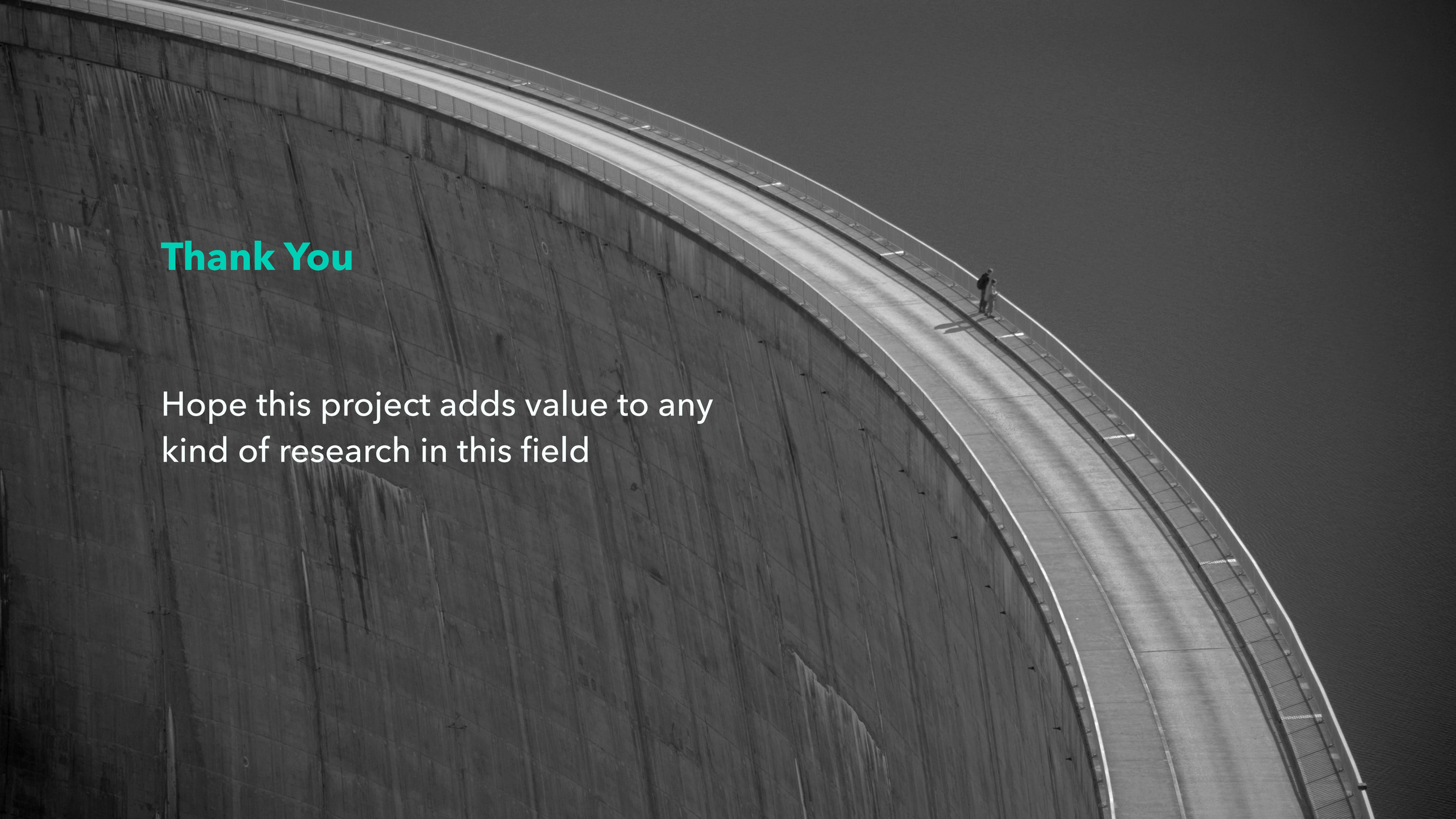
  2. At the instant of the FRB, AstroSat was pointing at (RA = 189.2, DEC = 62.3): about 157 degrees away from the nominal FRB direction.

PERFORMANCE ON EXAMPLES FROM PROBLEM DESCRIPTION

- 1. http://www.astronomerstelegram.org/?read=13354:
  Output is {(' 60.43819', ' 21.17461'), '0358+210', '0401+2110', '0401.7+2112', '040146+2110'}
- 2. http://www.astronomerstelegram.org/?read=13361:
  Output is {'060000.76-310027.83', '2017-84089.'}
  The second match is a false positive.
- 3. http://www.astronomerstelegram.org/?read=13351:
  Output is {'5000-9400'}
  This is a false positive.
- 4. http://www.astronomerstelegram.org/?read=13347:
  Output is {}

FUTURE WORK

- By generating more data, the classifier A can be improved significantly
- Extracting coordinates and/or other relevant information from tables.
- There is another category of coordinates in which an event can be represented in, called Galactic Coordinates. This can be extracted using the same principles used in this project.
- Important information from ATels, other that coordinates, like object name, error range can be extracted as per requirement.
- Deep Learning models like (NER) can be used if larger amount of data is provided.



Thank You

Hope this project adds value to any kind of research in this field