FINAL PROJECT REPORT

∗

**TEAM NEWS DIGEST**

Sravanthi Adibhatla†  
 Department of Computer Science  
 University at Buffalo  
 Buffalo New York USA  
 sadibhat@buffalo.edu

FirstName Surname  
 Department Name  
 Institution/University Name  
 City State Country  
 [email@email.com](mailto:email@email.com)

FirstName Surname  
 Department Name  
 Institution/University Name  
 City State Country  
 email@email.com

ABSTRACT

As elections are near approaching, there will be numerous events happening in countries like India, Indonesia and Thailand. We are focusing on the events, which will bring social unrest in the society such as Riots or Protests, which can possibly influence the elections. Since, the lengthy news articles are not very engaging enough, people lost the habit of news reading or following up with the current affairs. Our goal is to provide hand curated, short and crisp news preserving the original content. We are also providing tags such as event\_location, event\_date, parties/organizations involved so that it becomes easy for the end user to pick relevant content.

1.DATA SOURCES

We have used NewsApi, WebHose, and NYTimes to collect data. From the above sources mentioned, we are trying to extract URL from the json, which we get after hitting the API. Once we have the access to the URL, we will web scrape the URL to perform extraction and summarization.

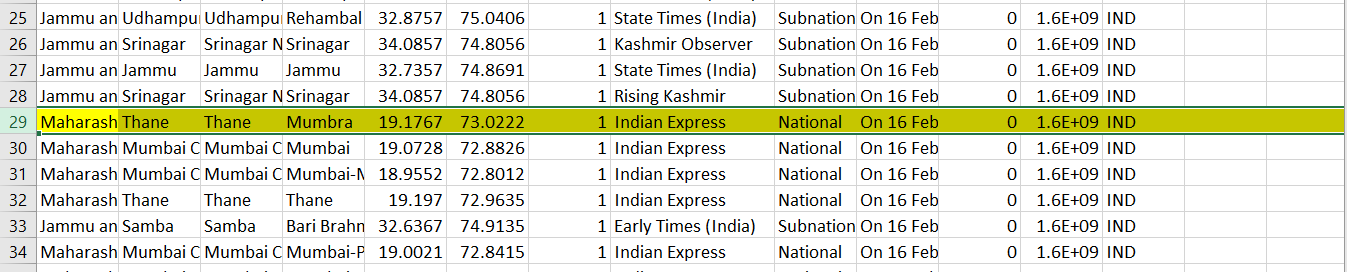
2. BASELINE

We are using ACLED as a benchmark to validate our model results. Randomly, we will pick few events from ACLED, and find the respective event’s URL online.

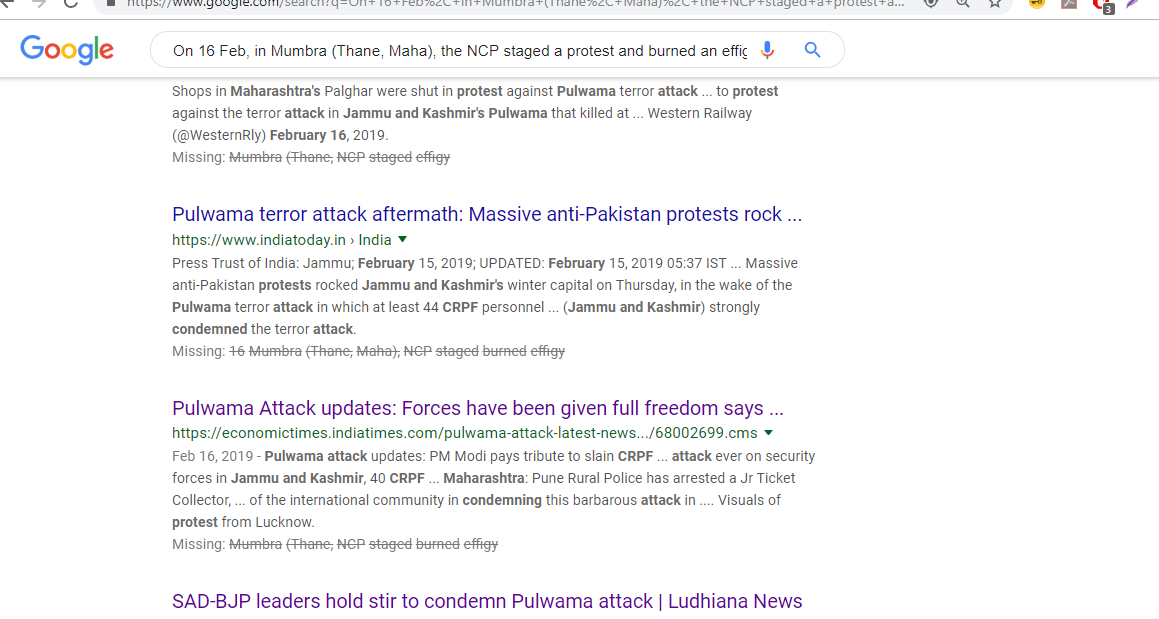
The event URL is send as input to the summarize function of the library we plan to use and in parallel the same URL is sent to extraction function which implements NER and tags the required fields like event\_date, location, parties/organization involved etc.

Later, we will compare the results with the ACLED attributes and summary using Evaluation Algorithm, which gives us the accuracy of the baseline model.

Let us look at an example that illustrates the above-mentioned steps:

Below is a data row from dataset that collected from ACLED, which is about recent Pulwama Attack. 

ACLED provided us that the data taken from Indian Express. If there is URL to the article is present, that would be the input to the system, but since that is not present, we have to search manually and find the article for the relative data row, which is present in ACLED.

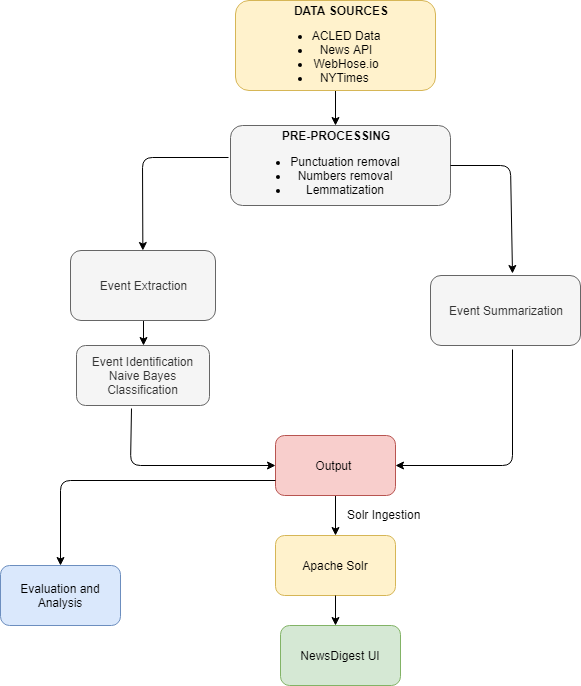


There are several URL’s which have the similar news article published, we picked article of Business Today because its URL is openly available for accessing the article.

<https://www.businesstoday.in/top-story/pulwama-terror-attack-live-updates-modi-govt-to-discuss-further-plan-amid-calls-for-strong-retaliatory-action/story/319258.html>

This is URL is the input to our extractor as well as summarizer, and the output summary as well as the attributes values are compared against ACLED standard using evaluation metrics.

3. SOLUTION ARCHITECTURE



The solution architecture depicts how the system works. We are performing extraction and summarization along with event identification. Then, the consolidated resultant json is ingested in Apache Solr and we created a simple search UI for our News Digest to show the results. In addition to that, we are evaluating our results with baseline using Naïve Bayes for Event Identification, Rouge for Summary and rest of the fields we assigned a score table and are calculating the mean for each slots like location, data and organization involved.

**4. EVENT EXTRACTION**

The task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. In most of the cases, this activity concerns processing human language texts by means of natural language processing (NLP). Here, we are using Spacy Library to perform Name Entity Recognition(NER) and are able to extract slots like GPE, LOC, ORG, DATE and are applying few techniques to match the original content of the article w.r.t. ACLED.

**4.1 EVENT TYPE IDENTIFICATION:**

We are classifying our articles into three classes:

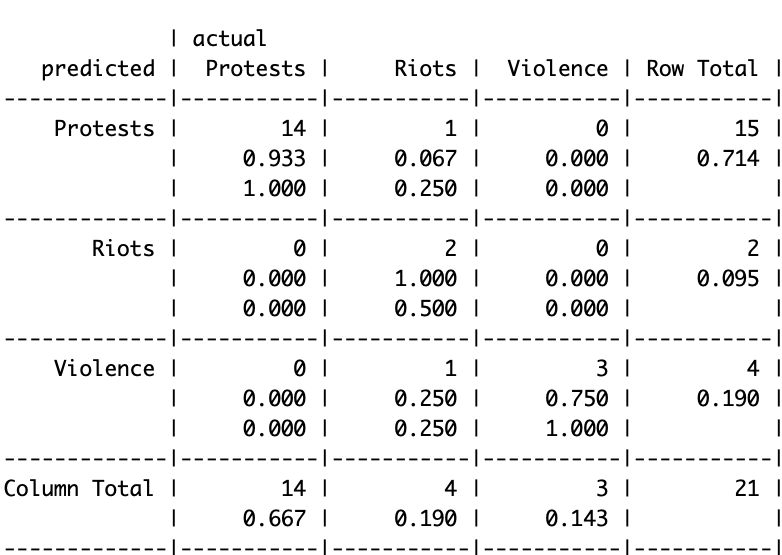
* Protests
* Riots
* Violence

We are using supervised learning algorithm Naïve Bayes for predicting event type from our input data.

Reasons for using Naïve Bayes algorithm is:

* Bayesian classifiers have been proven to be effective for Text classification, such as junk email (spam) filtering, author identification, or topic categorization
* Simple, fast, and very effective
* Does well with noisy and missing data
* Requires relatively few examples for training, but also works well with very large numbers of examples
* Assumes all features equally important and independent to each other.

We trained our model with 80 % of ACLED extracted articles and to test the model we have used remaining 20 %. We have achieved 89 percent accuracy. Below, is the confusion matrix, which gives clear picture of actual vs predicted results.



4.2 LOCATION EXTRACTION:

Every news article is in top-down approach. Therefore, the title of the article in most of the cases will definitely have location in it. We are performing NER on Title of the article, which we extract using Newspaper3k library.

If GPE is present in Title: We are performing Google Geo Coding to get the Address for the GPE present in the title.

If Title has no GPE tags: We iterate through all the GPE , and find the addresses length.

Ex: [Buffalo, California, Dallas]

Now, when we iterate and perform geocoding, the formatted addresses are:

* Buffalo, New York , USA
* California, USA
* Dallas, Texas , USA
* We are targeting at showing the most precise location, so whichever array size is 3 or more , we are considering that, here, Buffalo and Dallas.
* Since, it is a tie case; we keep a count of frequency of each GPE and check between Buffalo and Dallas, which has the highest count, and display that as the Event\_location.
* Suppose, if there are no three-sized array, we go for two- sized, i.e State, Country, and if no 2 sized, the just show the country.

4.3 DATE EXTRACTION:

We are using dateExtractor Library to extract published date from the article.

Then we are looking for special words like ‘today, yesterday, last Friday, on Saturday’ in the article text.

Then, we are associating each of these keywords with a number.

Finally, we are subtracting that number from the published date.

Ex:

Published date: 5/5/2019

Special Word: On Thursday

Event Date:

5th May = Sunday= number 7

7-4 = 3

So, 5th may – 3 days= 2nd may

5/2/2019 is Event Date

4.4 ORGANIAZTIONS/PARTIES EXTRACTION:

* We are doing NER tagging and then collected all the ORG tags present in the article text.
* We gave more weightage to the parties that are appearing in the title and summary.
* If we find ‘Some Group’ or ‘Group of ’, then, we are creating a new party with EVENT TYPE+ COUNTRY.

Ex: ‘Protestors (India)’, ‘Rioters (Indonesia)’

5. EVENT SUMMARIZATION

We are doing extractive summarization to find the summary of a news article. We are using NLTK Library, which uses text rank as the technique to find the summary.

Following are the steps to get the summary:

* Converting paragraphs into sentences
* Text processing
* Tokenizing sentences
* Find weight frequency of occurrence
* Replace words by weighted frequency in original sentences
* Sort sentences in descending order of Sum

We improved the summary accuracy by appending date to the summary like in ACLED.

Ex: On April 30, 19, + summary

Previously, we were getting top two sentences. Therefore, we experimented with different number of sentences and finally taking Top 4, which had better gist of content.

COMPARISION OF ACLED VS MODEL SUMMARY

Add a pic here or table

6. EVALUATION

6.1 SUMMARY EVALUATION:

We used ROUGE algorithm to evaluate the event summary. It uses precision, recall and f1 measure scores to calculate the accuracy. Below is the snippet of how ROUGE algorithm works:

Implementation of Rouge algorithm:

*Sample input:*

from PyRouge.pyrouge import Rouge r = Rouge()

acled\_summary = "The Kyrgyz President pushed through the law requiring the use of ink during the upcoming Parliamentary and Presidential elections In an effort to live up to its reputation in the 1990s as an island of democracy. The use of ink is one part of a general effort to show commitment towards more open elections. improper use of this type of ink can cause additional problems as the elections in Afghanistan showed. The use of ink and readers by itself is not a panacea for election ills."

model\_summmary = "The use of invisible ink and ultraviolet readers in the elections of the Kyrgyz Republic which is a small, mountainous state of the former Soviet republic, causing both worries and guarded optimism among different sectors of the population. Though the actual technology behind the ink is not complicated, the presence of ultraviolet light (of the kind used to verify money) causes the ink to glow with a neon yellow light. But, this use of the new technology has caused a lot of problems. "

[precision, recall, f\_score] = r.rouge\_l([acled\_summary], [model\_summmary])

print("Precision is :"+str(precision)+"\nRecall is :"+str(recall)+"\nF Score is :"+str(f\_score))

*Output Obtained:*

Precision is :0.446058091286

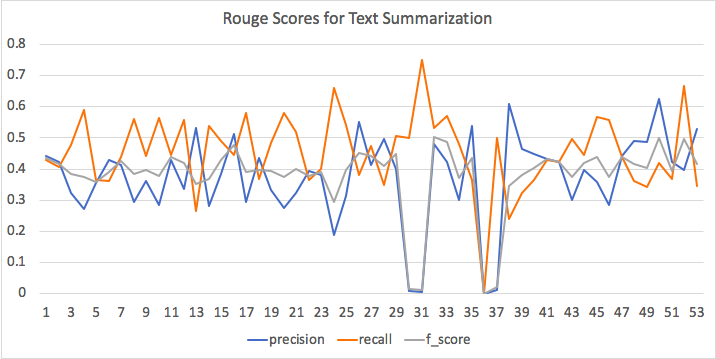
Recall is :0.439672801636

F Score is :0.442843380487

We have implemented this algorithm against our model\_summary with ACLED\_summary. Our data set contains 53 URLS and below is the Line graph, which shows how the values of all three variables vary.

Below is the line graph showing how the Rouge score values vary for each record:

*Baseline 1:*



6.2 LOCATION EVALUATION:

LOCATION EVALUATION:

In ACLED data, we have given following attributes for location in hierarchical order starting from specific location to country.

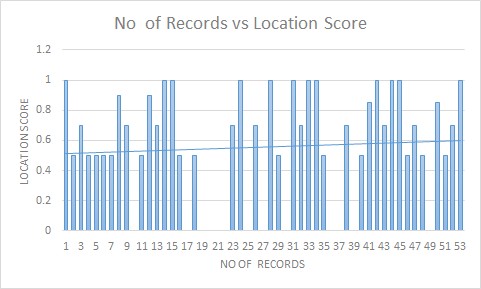
In our location evaluation strategy, we are planning to give weights to each attribute .Starting with higher weights to specific location to decreasing weights as we move up the hierarchical order.

|  |  |  |
| --- | --- | --- |
| Location Hierarchy | Location Values | Weights |
| Country | India | 0.5 |
| Admin1 | Uttar Pradesh | 0.7 |
| Admin2 | Azamgarh | 0.85 |
| Admin3 | Lalganj | 0.9 |
| Location | Nawapura | 1 |

Steps:

* We will start comparing the location obtained by our system with location hierarchy values given in the ACLED data. We will assign values according to the level of matching .
* So, here the first location is Model\_location, and remaining 4 fields are taken from ACLED. Since, in first row, we are able to fetch the most precise location, we are giving score as 1.
* In second row, if you observe, we were able to fetch country, which is not as precise as location so we are giving score as 0.5.

*Baseline 1:*



Overall score and accuracy of LOCATION is: 0.556 and 55.66 %

6.3 ORGANIZATIONS/PARTIES INVOLVED EVALUATION:

ACLED has parties involved in form of 4 different columns namely:

● Actor1

● Actor2

● Assoc\_actor1

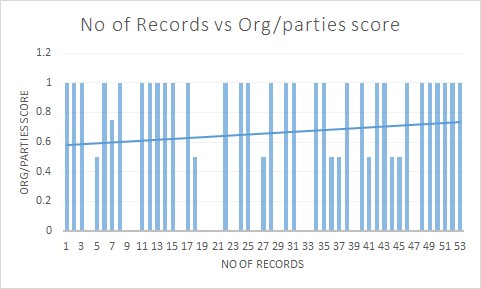
● Assoc\_actor2

We are combining all these and making a cumulative list which is ACLED’s Parties/Org involved.

Now, we are compare our generated list of parties/org to the ACLED’s list and if all the parties involved is present then its 100 % accuracy.

So, the formula is = no of matches / Total no of parties in ACLED

*Baseline 1:*

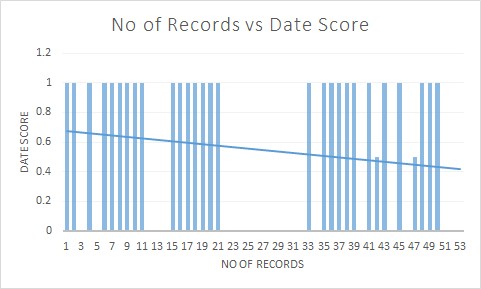


Overall score and accuracy of ORGANIZATIONS/PARTIES is: 0.6556 and 65.56%

6.4 DATE EVALUATION:

For the date evaluation, we are comparing the date obtained by our NER system to the date given in the ACLED gold reference date.

*Baseline 1:*



Overall cumulative score and accuracy of DATE is: 0.547 and 54.7%

COMPARISON OF BASELINE1 AND BASELINE 2

ACKNOWLEDGMENTS

Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

REFERENCES

[1] An Extensible Event Extraction System With Cross-Media Event By Fabio Petroni∗ , Natraj Raman∗, Timothy Nugent∗, Armineh Nourbakhsh†, Žarko Panić∗, Sameena Shah†, Jochen L. Leidner∗‡ Resolution

[2] Dhruv Gupta, Jannik Strötgen, and Klaus Berberich. 2016. EventMiner: Minings

Events from Annotated Documents. In Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval

[3] Jakub Piskorski, Hristo Tanev, Martin Atkinson, Eric Van Der Goot, and Vanni

Zavarella. 2011. Online news event extraction for global crisis surveillance. In

Transactions on computational collective intelligence V. Springer

[4] A survey on abstractive summarization techniques ,Pavan Kartheek Rachabathuni

[5] Event Detection from Text Data By Tomas Kala

[6]. Text Summarization Techniques: A Brief Survey by Krys Kochut, University of Georgia Athens

[7]. Design challenges and misconceptions in Named Entity Recognition by Lev Ratinov ,Dan Roth

[8]. Derczynski, Leon, et al. "Analysis of named entity recognition and linking for tweets." Information Processing & Management 51.2 (2015): 32-49.

[9]. A survey of named entity recognition and classification David Nadeau, Satoshi Sekine National Research Council Canada / New York University.

[10]. Text Summarization Techniques: A Brief Survey Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, Krys Kochut

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00