A Text Mining Approach for Forecasting Social Unrest and Violence in India

Ateendra Ramesh
ateendra@buffalo.edu
Graduate Student
University at Buffalo

Daniel Amirtharaj
damirtha@buffalo.edu
Graduate Student
University at Buffalo

Aashish Jain
ajain28@buffalo.edu
Graduate Student
University at Buffalo

ABSTRACT

Analysis and prediction of social unrest, riots and protests, is an active research topic that can help governments and the public avoid untoward situations that may cause damages to life and property. The primary objective of the system proposed in this paper is to forecast the probability of an outbreak of such events with a lead time of 2 days and within a window of 5 days. The data used for the aforementioned system is politically related Times of India headlines and tweets in chosen locations in India. The features employed in this paper are simple yet efficient in capturing location-specific trends. The average F1 score of the two models proposed in this paper 0.67 and 0.76 on the Twitter and Headlines datasets respectively; the performance of the models increase as the data generated by a location increases. Furthermore, in order to attribute these predictions with 2019 General Elections in India, the authors present a naive algorithm.

KEYWORDS

Datasets, Neural Networks, LSTM, Sentiment Analysis

ACM Reference Format:
Ateendra Ramesh, Daniel Amirtharaj, and Aashish Jain. 2019. A Text Mining Approach for Forecasting Social Unrest and Violence in India. In *Proceedings of Text and Web Mining (CSE635).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Social unrest, demonstration, protests and violence are a common occurrence in most developing countries, where governments tend to make quick policy changes, political parties are not well established or the county is not economically stable. The onset of these events can usually be attributed to other events that cause dissatisfaction among a significant group of people. Depending on the magnitude of such an event, the damage caused can be from marginal to significant loss to life and property. The prediction of these events in the near future, can prove useful to governments and appropriate law enforcement officials that can be prepared for the onset of such events.

Since, social unrest events are often caused by tensions that build up over time, the prediction of occurrence of such events before they occur can help in significantly controlling damages caused to life and property. Moreover, it can be observed from the ACLED dataset [9] that the occurrence of such events generally spike when a country is going through an election phase. With the increase in use of social media and most people sharing their opinions freely on such platforms, it is possible to study and analyze trends that may be indicative of future election related unrest situations.

2 RELATED WORK

There have been a lot of research work on the subject matter of predicting and analyzing social unrest via social media, however, a majority of these systems are focused on American state and national politics. Moreover, they rely on Twitter and news outlets such as the New York Times [1, 8].

In the system proposed by Mueller et al in [8], the authors make use of the New York Times in order to forecast political violence in middle-eastern countries. However, this system focuses on long-term prediction as opposed to short-term, which is necessary in the context of election-related violence.

In [2, 3], Blevins et al, developed a system in order to process tweets from gang-related youth from Chicago. Specifically, they focus on analyzing how conversations evolved into acts of violence, which is a common theme not just in gang violence but also in riots and protests. They extracted emotion scores for each tweet, of which emojis and emoticons were also given significance. The model proposed in this paper also focuses on the latter, by using Vader [5].

The model proposed by Bahrami et al in [1] uses Twitter in order to forecast a riot in the context of the 2016 US Presidential Elections. They used features of tweets collected for a particular day and extracted their polarity values as well state-specific features such as political leaning and affiliation. This research work proved to be extremely helpful in developing the model presented in this paper.

3 PROBLEM STATEMENT

The increase in availability of social media and news articles along with improvements in machine learning methods can be leveraged to make predictions about occurrence of social unrest and violence. Prediction of these events that are likely to occur in the near future, can prove useful to governments and appropriate law enforcement officials that can be prepared for the onset of such an event and minimize the damage to life and property.

The authors propose a system that uses headlines and tweets from Twitter, News Headlines of India from Kaggle and Times of India to predict the probability of social unrest or violence with a lead time of 2 to 7 days for all major cities listed in the ACLED dataset, trained using ACLED data as the ground truth.

CSE635, Spring 19, University at Buffalo
© 2019
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

4 DATA

This section explains the use of all datasets that are were explored or currently being used for feature extraction and prediction.

4.1 ACLED

The Armed Conflict Location & Event Data Project [9] (ACLED) is a disaggregated conflict analysis and crisis mapping project. ACLED is the highest quality, most widely used, real-time data and analysis source on political violence and protest in the developing world. Practitioners, researchers and governments depend on ACLED for the latest reliable information on current conflict and disorder patterns.

This dataset is used, both for training the prediction model as well as evaluating its performance. The first 75% share of events (in terms of time) that fall in the "Riots/Protests" and "Violence against citizens" categories are used as training labels, that are predicted, using training data extracted from other datasets. The trained model is then evaluated on the remaining 25% share of events.

4.2 News Headlines Of India Dataset

News Headlines of India Dataset [6] is a publicly available dataset on Kaggle that consists of about 2.9 million headlines from the newspaper, *Times of India*, spanning from 2001 to 2018 (inclusive). The headlines give an apt and short summary of the article, and serve as historical data. Headlines related to election violence and unrest were filtered and used to train the predictive model. Only data from the start of 2016 was used since ACLED (which acts as ground-truth), contains data only from 2016. Figure 1 shows a bar plot of the distribution of articles for different cities.

4.3 Twitter

Twitter is a rich source of real-time causal data that is widely accessible to most people. Politicians, rioters and protesters may engage in conversations or make statements that may be indicative of an event of interest. As mentioned in [1], understanding the frequency of tweets with negative sentiments at a certain instance can be useful in predicting future unrest events.

Here, tweets were collected using the Twitter API and the twython library using relevant keywords that are indicative of tweets that may be sent before the onset of an election related unrest event. This dataset spans from the end of February to this date, and more than 500,000+ tweets have been collected thus far, across locations present in the ACLED dataset.

4.4 The Times of India News Articles

News articles are the best source of information for any event that has occurred in the past. Unlike tweets and headlines, news articles provide an unbiased, holistic view of the situation and detailed information about events, and can be useful in prediction of social unrest events. Moreover, due to the limitation of the number of old tweets that could be accessed, the number of data points that could be used for training using ACLED as the ground truth was really small. Hence, the past news articles were used. About 80000 news articles were scraped from the *Times of India* website starting from 1-Jan-2016 until 5-May-2019.

5 DATA PRE-PROCESSING

5.1 News Headlines of India Dataset

The dataset consists of headlines encompassing a wide range of topics, and hence was filtered using a rule-based filter, which removed all headlines not consisting relevant keywords such as BJP, riots, protests, elections etc. No machine learning model was used for this task as the number of words in a headline ranged from 5 to 10, and it seemed futile to do so. It was also observed that headlines typically had information about involved parties as well as the event(s).

This dataset consists of the publishing date, category of headline and headline, which are used to infer the location of interest. The location information is furthermore used to group the filtered headlines based on location. The filtered headlines with locations present in ACLED are considered, and the rest eliminated. About 33,000 data points were obtained, which were used for feature extraction.

6 FEATURE EXTRACTION

This section elaborates the features extracted for all the datasets mentioned above.

6.1 Location-wise grouping of tweets

It was observed that most of the tweets collected had no information about the exact location from where the tweet was made. This is due to the fact that most users turn-off location sharing while making the tweet. Location is one of the important information that every tweet should have, but the number of people who share their exact location at the time of tweeting is low and such tweets might not be enough to make predictions. Hence, the tweets were attributed to locations that were mentioned in them. If a specific location was not given, e.g. a State was given instead of a city, the tweet data point was added to the data points of all the cities of the state. This is important because in the case of Social unrest prediction, recall should be high, so that all events that might occur may be identified, and any damage to life or property that might occur if the protests turn violent, may be prevented.

6.2 VADER Sentiment Analyzer

The VADER (**V**alence **A**ware **D**ictionary and s**E**ntiment **R**easoner) [5] tool is a powerful sentiment analysis model that is built to perform on social media content as it emphasizes on the emoticons, slang, emojis and also capitalization of text, which might convey intensity. Given a single sentence, VADER returns a positive, negative, neutral and compounded sentiment, which are used as features for all documents.

6.3 Computing Features

The features mentioned below have been computed for each document (tweet/headline).

- **Number of documents** - This is a total number of documents acquired on each day.
- **Sentiment features** - All documents acquired for that particular day are analyzed through VADER and their collective positive, negative, neutral and compound sentiment values are averaged as 4 different features.

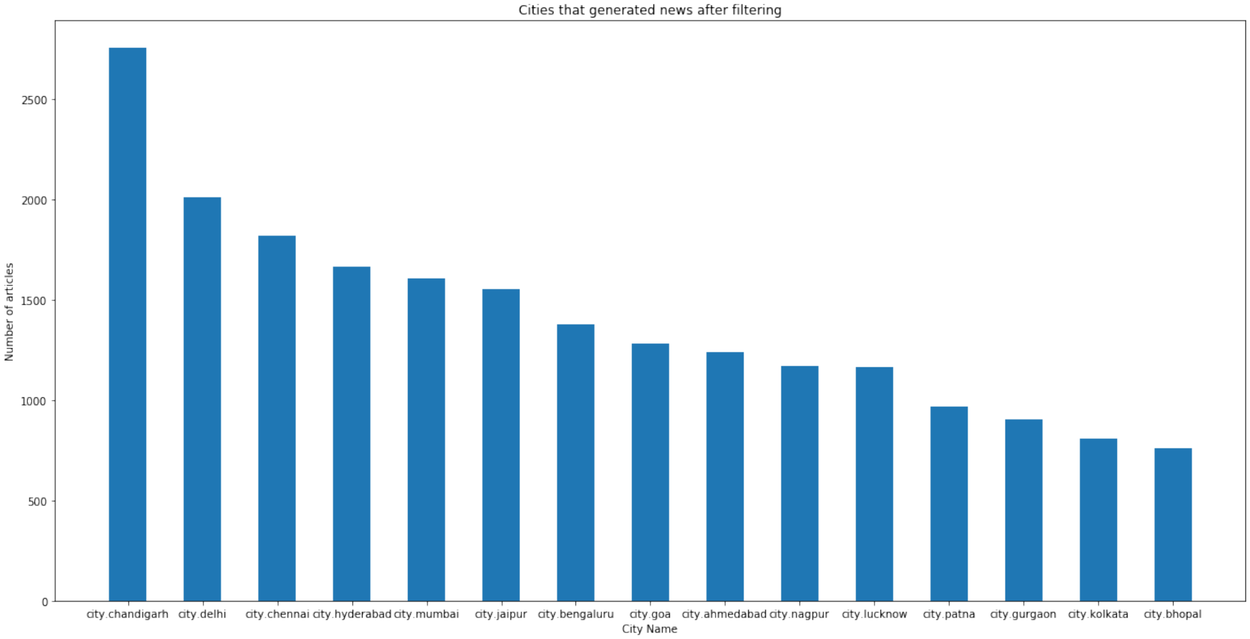

Figure 1: Cities that generated most news in Headlines dataset

- **Number of negative documents** - From the sentiment values extracted, the number of negative documents were also used as features as it proved to be a key factor in determining an riot outbreak in [1] by Bahrami et al.

Thus, a total of 6 features were used for each day.

6.4 Labels

Models for specific locations were trained to produce an output of whether an event has occurred within a window of 5 days (i+2 to i+7), post a lead time of 2 days from a starting date i (features for date i is used). This prediction was done by first training the model over a sliding window (by incrementing i) over a portion of dates used exclusively for training (first 75%). The model was then used to label (predicted label) the occurrence of events on the remaining testing dates (remaining 25%). The occurrence of an event was labelled (actual label) based on whether an event had occurred in the 5 day window in the ACLED dataset.

7 SOLUTION ARCHITECTURE

The current solution architecture is shown in Figure 2. For the tweets and news headline dataset, posts and news articles from various sources are first filtered based on keywords. These filtered keywords are then grouped by location for which predictions are to be made. Features are calculated for the datasets as shown in Section 6. Using, historical data we have trained different ML (classic and deep learning) model(s) that would be able to make a prediction for occurrence of events, given real-time data (present day) in the next 2 to 7 day window.

8 MODELS

Currently, classical machine learning models such as Extremely Randomized trees (Extra trees/ET), Logistic Regression; of which the ET model seemed to work best for most cities. Deep learning techniques such as 1D Convolution and LSTM models have also been used for making predictions, but they fail to outperform ET which the authors attribute to fewer data points. Currently, a model is deployed for each of the locations in each dataset.

9 RESULTS

In order to evaluate the performance of the aforementioned ML models, 4 metrics were used as key performance indicators - Test Accuracy, Precision, Recall and F1-Score. The first metric is to get an idea of general performance whereas the rest account for the model's performance on skewed data.

9.1 Headlines Dataset

This dataset as mentioned in Section 4.2, has the most overlap with the ACLED dataset, the above metrics have been applied for the entirety of the dataset as well as a continual prediction simulation. After extracting features, the data is sorted by publish date and is ***not randomly permuted***, i.e., the temporal ordering is maintained. The first 75% of the data was used for training, and the rest for testing. Table 1 showcases the results for the ET model's performance over the set of locations for which the headlines dataset was generated.

9.1.1 Continual Predictive Analysis. The test dataset (sorted by publish date) was partitioned into batches of size 10, for a batch b_i, a model was trained on batches $\{b_0 \ldots b_{i-1}\}$ and tested on b_i and so on. For this task, ET models were used as it was empirically

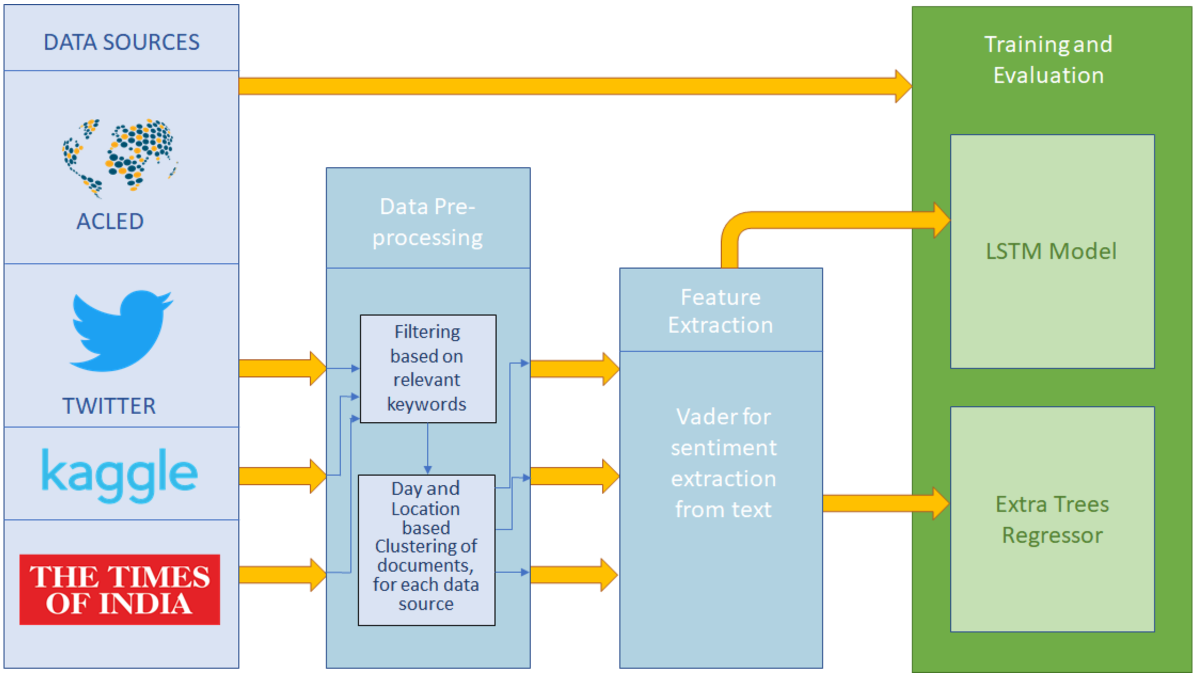
DATA SOURCES
ACLED
TWITTER
kaggle
THE TIMES OF INDIA
Data Pre-processing
Filtering based on relevant keywords
Day and Location based Clustering of documents, for each data source
Feature Extraction
Vader for sentiment extraction from text
Training and Evaluation
LSTM Model
Extra Trees Regressor

Figure 2: Current Solution Architecture

observed to result in the best performance under this setting as well. Some interesting observations were that cities such as Jammu and Srinagar had extremely good performance, but that is due to the nature of conflicts in these cities as opposed to the model's efficiency. = The city of Chandigarh generated the most news headlines in the dataset's timeline, Figure 3 shows the performance of the model over batches wherein the model has 2 distinct points of dip in performance and also recovers promptly.

Another city that exhibited similar patterns but did not generated as much news as Chandigarh is Kolkata as can be seen in Figure 1. The continual performance of Kolkata can be seen in Figure 4.

9.2 Twitter Results

The same temporal train-test split mentioned in Section 9.1 was used in partitioning this dataset as well. Twitter data as stated in Section 4.3 consists of about 500,000 documents in and of itself for a short amount of time. Table 2 consists of the efficiency of the model over all locations in the dataset. However, continual analysis was not performed owing to a shortage in the data available.

A general but trivial trend that was observed was that cities that generated the most tweets such as New Delhi, Jammu and SriNagar outperformed those locations with fewer data points. It was also seen that recall over most locations remained consistently high, which is a desirable property of a good predictive model.

9.3 Continual & Progressive Learning

Continual learning is an important part of the machine learning models in general and especially in sequential data wherein an ML model going "stale", and becoming a victim of concept drift is highly likely event. The technique mentioned in Section 9.1.1 is a naive and straight-forward implementation but clearly can be improved by choosing specific data points to train on as opposed to training on all the data. The authors are currently researching methods mentioned by Gama et al in [4] in order to better understand state of the art techniques and incorporate them into the existing system.

9.4 Correlation with elections

In order to correlate the riots/protests with election, we used the percentage of documents that can be placed in the "election" topic, via classification based on rule-based models. If the event classified to be in "election" topic has a strong probability of occurrence, then we can state that the event that might occur is related to election. The classification was observed to not perform well which could be attributed to the informal language used in tweets and huge number of tweets.

The authors also trained a Doc2Vec model on about 50,000 articles scrapped from The Times of India website with the idea of clustering similar documents and using information of these similar documents to provide context in making predictions but the data used wasn't enough for the model to learn relationships between the documents. Also, the model would require periodic re-training to get the most recent related documents as new documents would have to be included. However, training Doc2Vec also trains a Word2Vec model alongside. The authors observed that it learned few relationships well for e.g., the two closest neighbours of Modi were Nehru and Vajpayee and attempted to use this to construct bag of words model for categorizing the event as Election-related and not Election related. However, the results were not promising as not all relationships were learned due to the limited data fed to the model and the former approach outperformed the latter.

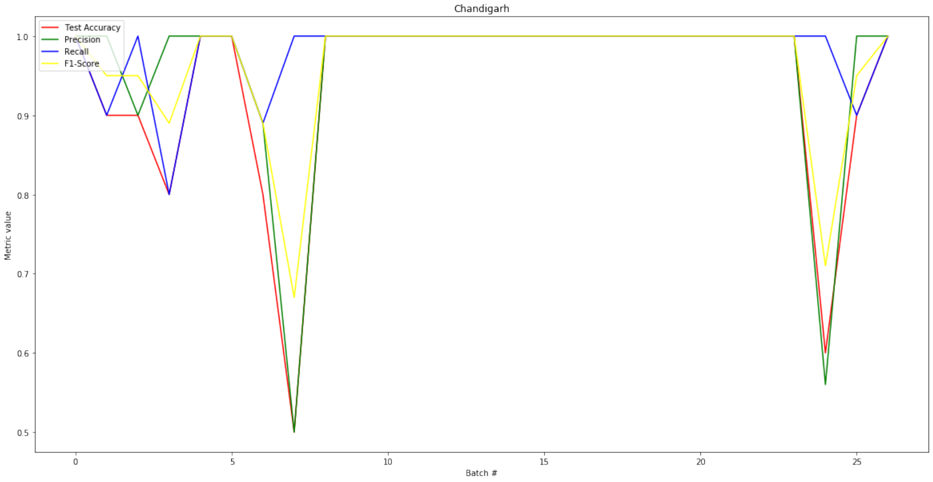

Figure 3: Chandigarh continual predictions - Headlines Dataset

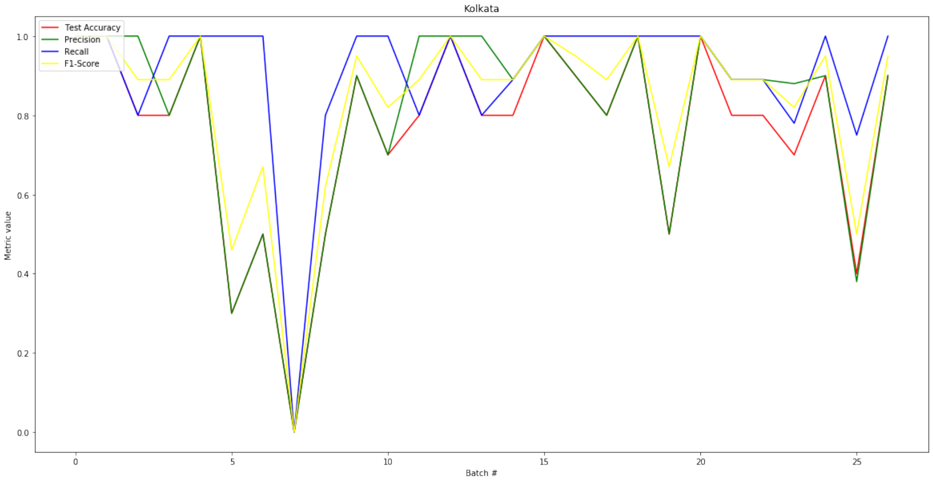

Figure 4: Kolkata continual predictions - Headlines Dataset

10 FUTURE WORK

Doc2Vec is an adaptation of Word2Vec that generates vectors for sentence(s), paragraph(s) and/or document(s) [7]. The authors plan to use it for clustering similar documents (news article or tweet) to provide context to the query document which could be used to make predictions. Unlike, the current model which just uses real-time sentiments extracted via tweets and news headlines, this would help in attributing the cause of any protest or riot to a given event. Moreover, a query document will likely exhibit behavior similar to its neighbours which could be useful while making the predictions.

11 CONCLUSION

Thus, the authors have proposed a social unrest forecasting engine using a set of simple and efficient features powering 2 models built on Twitter and headlines data by achieving F1 scores of 0.67 and 0.76 respectively. The authors also observed, for a particular location, there was a positive correlation between a model's efficient functioning with the amount of data the location produces.

ACKNOWLEDGMENTS

The authors would like to thank our Professor Rohini K. Srihari and our TA Lu Meng for guiding us with theoretical and practical advice and feedback throughout the semester while working on this project by setting milestones for tracking progress.

REFERENCES

[1] Mohsen Bahrami, Yasin Findik, Burçin Bozkaya, and Selim Balcisoy. 2018. Twitter Reveals: Using Twitter Analytics to Predict Public Protests. *CoRR* abs/1805.00358 (2018). arXiv:1805.00358 http://arxiv.org/abs/1805.00358

[2] Terra Blevins. 2016. Natural Language Processing Applications for Prediction of Violence in Gang-Related Social Media.

[3] Terra Blevins, Robert Kwiatkowski, Jamie Macbeth, Kathleen McKeown, Desmond Patton, and Owen Rambow. 2016. Automatically processing tweets from gang-involved youth: towards detecting loss and aggression. In *Proceedings of COLING*

Table 1: Performance of Augmented Headlines model trained from Jan-2016 to Dec-2017 and tested on Jan-19 till April-19

Location	Test Accuracy	Precision	Recall	F1 Score
Agartala	0.37	0.67	0.05	0.09
Ahmedabad	0.53	0.51	0.37	0.43
Amritsar	0.95	0.96	0.99	0.97
Bathinda	0.98	0.98	1	0.99
Bengaluru	0.71	0.88	0.77	0.82
Bhubaneswar	0.76	0.81	0.91	0.86
Chandigarh	0.95	0.96	0.99	0.97
Chennai	0.72	0.79	0.88	0.83
Coimbatore	0.64	0.68	0.88	0.77
Dehradun	0.6	0.6	0.97	0.74
Guwahati	0.79	0.86	0.9	0.88
Hyderabad	0.48	0.8	0.51	0.62
Imphal	0.87	0.89	0.98	0.93
Jaipur	0.48	0.47	0.71	0.56
Jalandhar	0.88	0.9	0.98	0.94
Jammu	1	1	1	1
Karnal	0.53	0.53	1	0.7
Kolkata	0.84	0.9	0.92	0.91
Lucknow	0.71	0.73	0.96	0.83
Ludhiana	0.93	0.96	0.97	0.96
Madurai	0.46	0.59	0.23	0.33
Patiala	0.8	0.87	0.9	0.88
Patna	0.46	0.73	0.43	0.54
Puducherry	0.44	0.43	0.95	0.6
Ranchi	0.62	0.64	0.94	0.76
Sangrur	0.61	0.61	1	0.76
Shimla	0.42	0.4	0.93	0.56
Srinagar	1	1	1	1
Thiruvananthapuram	0.68	0.74	0.88	0.8
Average	0.7	0.75	0.83	0.76

2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2196–2206.

[4] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM computing surveys (CSUR)* 46, 4 (2014), 44.

[5] Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media.*

[6] Rohit Kulkarni. [n. d.]. News Headlines of India 2001-2018 [CSV data file]. ([n. d.]). https://doi.org/10.7910/DVN/J7BYRX

[7] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. *CoRR* abs/1405.4053 (2014). arXiv:1405.4053 http://arxiv.org/abs/1405.4053

[8] Hannes Mueller and Christopher Rauh. 2018. Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review* 112, 2 (2018), 358–375.

[9] Clionadh Raleigh, Andrew Linke, HÃěvard Hegre, and Joakim Karlsen. 2010. Introducing ACLED: An Armed Conflict Location and Event Dataset: Special Data Feature. *Journal of Peace Research* 47, 5 (2010), 651–660. https://doi.org/10.1177/0022343310378914 arXiv:https://doi.org/10.1177/0022343310378914

Table 2: Performance of Twitter data using Extra-Trees model

Location	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	+'s in data
Agartala	1	0.19	0.94	0.19	0.23	0.63
Ahmedabad	1	0.75	0.75	0.75	0.75	0.38
Amritsar	1	0.75	0.75	1	0.86	0.94
Bathinda	1	0.75	0.87	0.75	0.8	0.9
Bengaluru	1	0.94	0.94	1	0.97	0.9
Bhubaneswar	0.96	0.75	0.65	0.75	0.7	0.83
Chandigarh	1	0.94	0.94	1	0.97	0.9
Chennai	1	0.75	0.75	1	0.86	0.89
Coimbatore	1	0.5	0.81	0.5	0.47	0.68
Dehradun	1	0.94	1	0.94	0.97	0.19
Delhi-New Delhi	1	0.94	0.94	1	0.97	0.97
Gurgaon	1	0.62	0.59	0.62	0.56	0.25
Guwahati	1	0.44	0.62	0.44	0.42	0.71
Hyderabad	1	0.88	0.88	1	0.93	0.9
Imphal	0.98	0.75	0.71	0.75	0.72	0.71
Jaipur	0.91	0.38	0.38	0.38	0.38	0.33
Jalandhar	1	0.75	0.87	0.75	0.8	0.78
Jammu	1	0.94	0.94	1	0.97	0.98
Karnal	1	0.25	0.08	0.25	0.12	0.52
Kolkata	1	0.69	0.69	1	0.81	0.92
Lucknow	1	0.19	0.59	0.19	0.22	0.52
Ludhiana	1	0.94	0.94	1	0.97	0.95
Madurai	1	0.44	0.44	1	0.61	0.46
Patiala	1	0.69	0.74	0.69	0.71	0.79
Patna	0.96	0.56	0.56	0.56	0.56	0.49
Puducherry	1	0.31	0.34	0.31	0.33	0.6
Pulwama	1	0.5	0.66	0.5	0.5	0.41
Ranchi	0.89	0.88	1	0.88	0.93	0.29
Salem	1	0.31	0.64	0.31	0.34	0.49
Sangrur	1	0.56	0.9	0.56	0.63	0.68
Shimla	0.83	0.44	0.46	0.44	0.4	0.41
Srinagar	1	0.94	0.94	1	0.97	0.98
Thiruvananthapuram	1	0.5	0.46	0.5	0.47	0.76
Tiruchirappalli	1	0.69	1	0.69	0.81	0.38
Average	0.99	0.64	0.73	0.7	0.67	0.66

Table 3: Twitter results with LSTM

Location	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	+'s in data
Agartala	0.95	0.5	1	0.5	0.67	0.58
Ahmedabad	0.68	0.64	0.64	1	0.78	0.35
Amritsar	1	0.71	0.71	1	0.83	0.93
Bathinda	0.88	0.93	0.93	1	0.96	0.89
Bengaluru	0.93	0.93	0.93	1	0.96	0.93
Bhubaneswar	0.93	0.79	0.79	1	0.88	0.8
Chandigarh	0.9	0.93	0.93	1	0.96	0.91
Chennai	0.93	0.71	0.71	1	0.83	0.87
Coimbatore	0.78	0.21	0.21	1	0.35	0.64
Dehradun	0.95	1	1	1	1	0.07
Delhi-New Delhi	0.98	0.93	0.93	1	0.96	0.96
Gurgaon	0.15	0.43	0.43	1	0.6	0.22
Guwahati	0.78	0.36	0.36	1	0.53	0.67
Hyderabad	0.9	0.86	0.86	1	0.92	0.89
Imphal	0.83	0.29	0.56	0.29	0.29	0.67
Jaipur	0.93	0.43	0.47	0.43	0.44	0.36
Jalandhar	0.85	0.93	0.93	1	0.96	0.87
Jammu	1	0.93	0.93	1	0.96	0.98
Karnal	0.68	0.57	0.67	0.57	0.57	0.45
Kolkata	0.98	0.71	0.71	1	0.83	0.91
Lucknow	0.63	0	0	0	0	0.47
Ludhiana	0.95	0.93	0.93	1	0.96	0.95
Madurai	0.71	0.36	0.36	1	0.53	0.38
Patiala	0.73	0.86	0.86	1	0.92	0.76
Patna	0.85	0.5	0.5	1	0.67	0.47
Puducherry	1	0.57	0.57	0.57	0.57	0.55
Pulwama	0.93	0.64	0.64	1	0.78	0.33
Ranchi	1	0.93	1	0.93	0.96	0.18
Salem	0.54	0.07	0.07	1	0.13	0.42
Sangrur	0.78	0.14	0.14	1	0.25	0.64
Shimla	0.88	0.71	0.73	0.71	0.71	0.45
Srinagar	1	0.93	0.93	1	0.96	0.98
Thiruvananthapuram	0.76	0.64	0.64	1	0.78	0.73
Tiruchirappalli	0.71	1	1	1	1	0.36
Average	0.84	0.65	0.68	0.88	0.72	0.64