Monitoring and Predicting Social Unrest: Baseline Approach

Sanyam Gupta
University at Buffalo
sanyamgu@buffalo.edu

Aditya Jagtap
University at Buffalo
ajagtap@buffalo.edu

Prateeksha Singh
University at Buffalo
psingh44@buffalo.edu

1 Abstract

.This project consists of analyzing social unrest data and predicting protests. It comprises of three tasks: Information Extraction, Summarization and Event Prediction. We try to solve these using transformer based approaches for the most part, while comparing our results with the baselines obtained with more traditional models like CNNs, BiLSTMS and Sequence to Sequence Transformer. The Information Extraction task follows a multiple models approach to predict 8 target variables from a single sentence, with different models for different groups of columns: Multiclass classification for categorical variables, and NER for a numeric and text variable. For Summarization we have implemented a Sequence to Sequence Transformer model to generate summarized english text from a given set of feature keywords. For Event prediction we have contrasted the performance of BiLSTM and BERT. Although the features used with each were different with those for transformer being vastly information rich. Moreover the task involves using News data, since BERT is pre-trained on NEWS data, using this specific transformer was a straightforward decision. The Transformer model is inherently more sophisticated so is the improvement in results in not surprising.

2 Introduction

Social unrest is hypothesized to exhibit patterns, which if detected early can facilitate mitigation efforts. This task consists of 3 subtasks as follows:

- Information Extraction: Extracting the number of fatalities, event type, sub event type, actor 1, inter 1, actor 2, inter 2, interaction and location from the summary of reported events. Our solution proposes to extract the categorical variables using multi-class classification with a BERT-based transformer, leveraging the dependencies between certain target variables. The non categorical variables are extracted using NER.
- Summarization: Generating summarized text from given list of ACLED data points such as event date, event-type, actors, interactions etc. We decided to look at this task as a language translation problem where we need to translate a language spoken only in feature keywords to english language.
- Event Prediction: Predicting the number of Protests on a particular day. This problem is formulated as a regression problem. Using data from previous days the model predicts the number of protests on a future date. Similar approach was adopted in (Srihari, 2019)

3 Literature Survey

3.1 Information Extraction

We primarily studies the approaches , starting with an overview of approaches for NER and Sentence classification based problems (Nadeau, 2014) Overview, A survey of named entity recognition and classification

: https://nlp.cs.nyu.edu/sekine/papers/li07.pdf
Baselines model to

Bidirectional LSTM-CRF for Named Entity Recognition: https://aclanthology.org/Y18-1061.pdf and CNN

ideas from

Jay Alammar: https://jalammar.github.io/illustrated-word2vec/

and transformers

resort to for sentence classification using a BERT-based model (mccormick, 2019)

https://keras.io/examples/nlp/ner$_{t}ransformers/$

3.2 Summarisation

For task 2 we considered it as an end to end language translation problem and we reviewed multiple approached and found out the one (Tianyu, 2017) suited best for the task since they were generating summary text from a similar tabular format.

3.3 Event Prediction

This task can be formulated as a time series prediction problem. This is a actively researched area of NLP. Using heterogeneous data sources in unrest prediction was demonstrated by (Meng, Srihari, 2019). They also proposed how was can predict far into future in another study of theirs (Meng, Srihari, 2019). The baseline model was formulated as a single step uni variate time series prediction problem. The applicability of LSTM for this problem domain is discussed in (Zang et al, 2018). Transformers were introduced by (Kaiser et al, 2017). They have proved to be a game changer due to their applications in NLP. BERT is a type of transformer introduced by (Toutanova et al, 2019). It has been pre-trained on News corpus.

4 Methods and Model architectures

4.1 Information Extraction

The dataset comprises of 1 features column (notes) and 11 target variables.

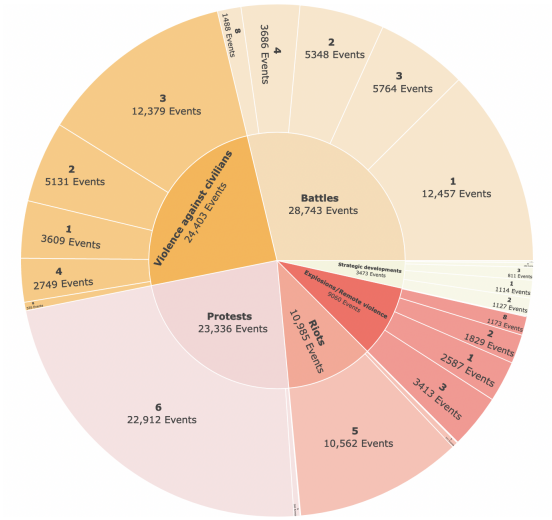
4.1.1 Field Types, Problem Type

- Fatalities: **Text/numeric**; Location: **Text**
- Event type, Sub event type, Inter 1, Inter 2: **Categorical**; Interaction: Derived (can be computed and hence dropped from training)
- Actor 1, Actor 2: **Categorical + Text**

	NOTES	EVENT_DATE	SOURCE	FATALITIES	EVENT_TYPE	SUB_EVENT_TYPE	ACTOR1	INTER1	ACTOR2	INTER2	INTERACTION	LOCATION
0	Three people were killed while 27 others injur...	29-August-2012	Statesman (Pakistan)	3	Explosions/Remote violence	Remote explosive/landmine/IED	Unidentified Armed Group (Pakistan)	3	Civilians (Pakistan)	7	37	Jacobabad
1	Government security forces opened fire at a pr...	03-May-2014	Undisclosed Source	0	Violence against civilians	Attack	Military Forces of Somalia (2012-2017)	1	Civilians (Somalia)	7	17	Baidoa

Categorical Fields are limited in the possible values they can take and hence could be solved by a Multi-class Classification. The difference from a pure Machine learning CNN approach would be the embedding incorporating implicit meaning of sentences rather just a BoW model.

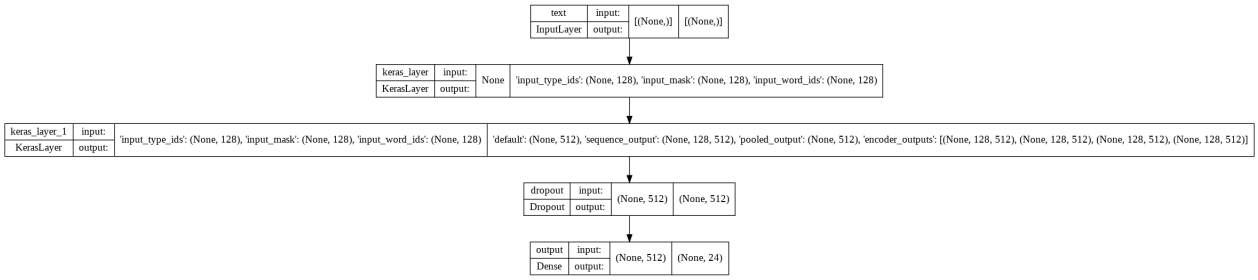
Assuming that Location and fatalities field values are usually explicitly present in the summary itself, we decided to follow an approach where using just the training set, sentences can be tagged by simple matching field value to closest matching word in sentence and used to train an NER model. Regression is also tried as viable method for fatalities.



4.1.2 NLP models

Transformers can account for an entire sentence in one go rather than piecemeal tokens, using Self Attention. BERT (Bidirectional Encoder Representations from Transformers) is a Machine Learning technique based on transformers, i.e. attention components able to learn contextual relations between words.

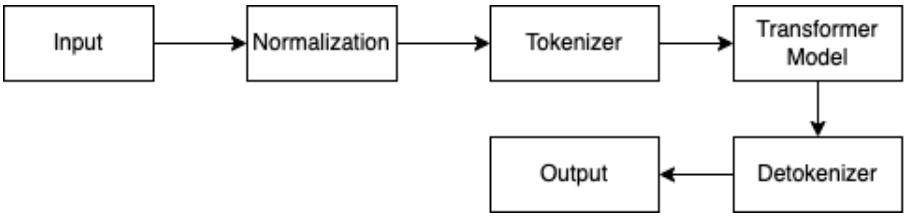
The task of classification refers to the prediction of a class for a given observation. For this reason, the only needed input to train such a model is a dataset composed of text samples, new report notes in our case. There are converted into BERT-based embeddings before passing to a transformer layer, with a dropout. Separate columns are the treated as the associated labels. In consolidated pairs of columns, every possible pair of values of those two columns is treated as a label and one-hot encoded as the the Dense output layer.



For NER, in order to use a Pre-trained BERT model, the notes have to be preprocessed and infused with the tags of location and fatalities fields.

4.2 Summarization

We have implemented a sequence to sequence network where we have used 5 encoder and decoder layers. Encoder is initially passed data which has already been tokenized and passes it through series of network layers each as multi-head attention and feedforward layers and and decoder receive it along with tensors which are later together trained in an end to end manner. We trained our model on 10% of the given dataset.



4.3 Event Prediction

ACLED data dump is used without augmenting it from external data sources.

4.3.1 Final

Data Pre-processing:
Out of the numerous columns we have used the EVENT_DATE, EVENT_TYPE and NOTES columns. For each EVENT_DATE, number of PROTESTS have been counted from EVENT_TYPE column and all NOTES have been concatenated for that day. The number of PROTESTS is the target variable and combined notes are used as the input. All the notes for a historical window are given as an input and we are predicting lead time days in the future.

History for: **3 days** Historical window

Predict **1 day** in future Lead Time

```
[55] temp
```

Combined EVENT_DATES in historical window

Protest counts on EVENT_DATE + Lead Time

Combined Notes

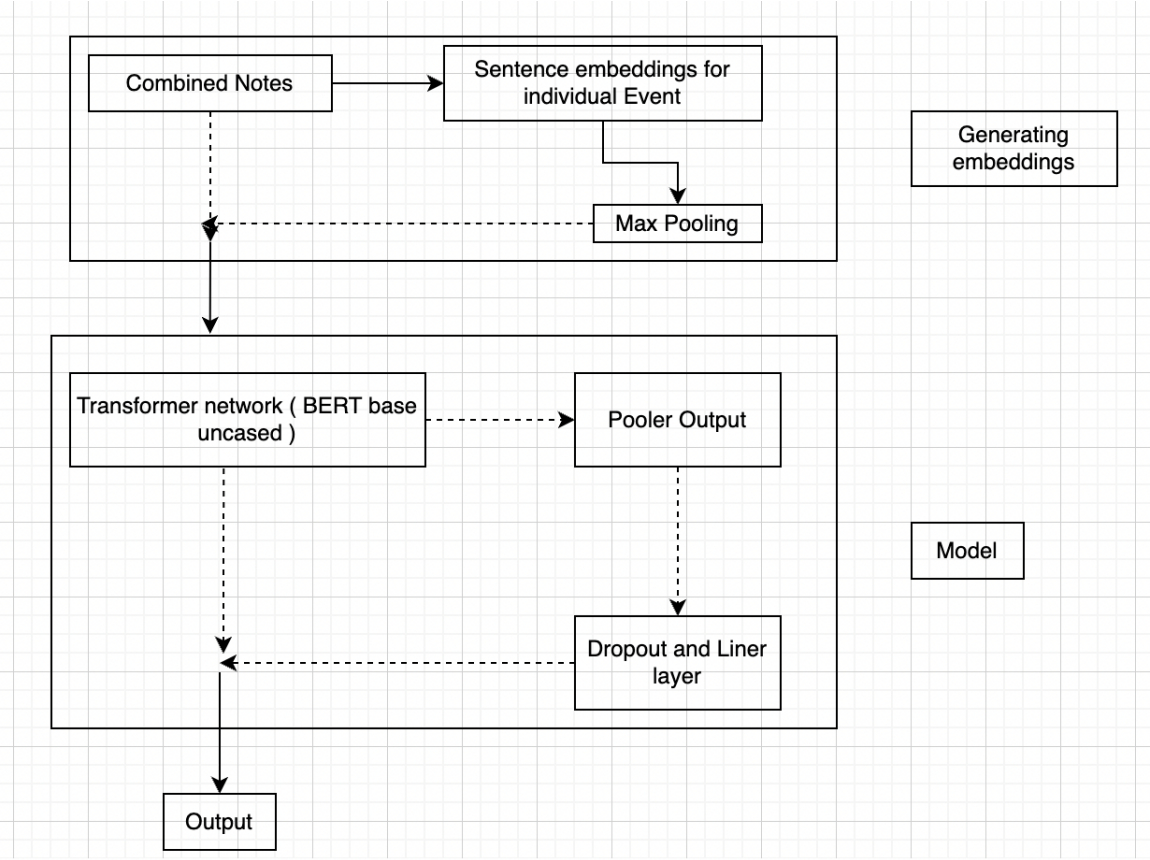
	EVENT_DATE	NOTES	PROTEST_COUNT
0	[2018-01-01, 2018-01-02, 2018-01-03]	[Members of the Safai Karamchari Union protest...	22
1	[2018-01-02, 2018-01-03, 2018-01-04]	[On 2 January 2018, a mentally unstable man wa...	30
2	[2018-01-03, 2018-01-04, 2018-01-05]	[An unidentified group of people hurled a gren...	14
3	[2018-01-04, 2018-01-05, 2018-01-06]	[On January 4, the Border Security Forces atta...	12
4	[2018-01-05, 2018-01-06, 2018-01-07]	[On January 5, in Puducherry, the Puducherry S...	58
...	...	...	...
1474	[2022-01-14, 2022-01-15, 2022-01-16]	[On 14 January 2022, 2 SSB and 1 BSF personnel...	86
1475	[2022-01-15, 2022-01-16, 2022-01-17]	[On 15 January 2022, dozens of journalists hel...	55
1476	[2022-01-16, 2022-01-17, 2022-01-18]	[On 16 January 2022, a group of Hindu donation...	45
1477	[2022-01-17, 2022-01-18, 2022-01-19]	[On 17 January 2022, around 20 people stopped ...	54
1478	[2022-01-18, 2022-01-19, 2022-01-20]	[On 18 January 2022, an armed clash occurred b...	37

1479 rows × 3 columns

Model:
The embedding for each individual NOTE for all the days are generated and max pooling is done over the resultant embeddings. The max length of each embedding is fixed to 200, so the resultant max pooled vector is also 200 dimensional. This is given as a input to BERT transformer. The transformer's pooler_output is an input to dropout and linear layer which given the predicted number.

	Milestone 2		Milestone 3		
	Method/Model	Balanced F1 score	Method/Model	Balanced F1 score	Accuracy
Event Type	Multiclass - CNN	0.3	Multiclass - Transformer	0.45	0.71
Sub Event Type	Multiclass - CNN	-	Multiclass - Transformer	0.10	0.61
Inter1	Multiclass - CNN	0.08	Multiclass - Transformer	0.07	0.50
Inter2	Multiclass - CNN	-	Multiclass - Transformer	0.06	0.56
Inter1 + Inter2			Multiclass - Transformer	0.20	0.41
Event Type + Sub Event Type			Multiclass - Transformer	0.02	0.25
Location, Fatalities	NER - BiLSTM	0.29	NER - Transformer	0.34	0.56

Table 1: Information Extraction: F1 scores



5 Results

5.1 Information Extraction

[Refer Table 1].

We observe that for categorical columns:

Trial 1: Single Columns for Event type, Sub Event Type, Inter1, Inter2 Results were found to be Average, 50-60% Accuracy, Although Testing F1 score was 0.45 for Event (easiest to predict), whereas for Inter was lowest at 0.06.

Trial 2 with Combining columns, dependent and independent pairings; Inter1 Inter2 gave Average results: 40% Accuracy, 0.20 Testing F1 score. But this gives us a single model to predict both columns. Event_type and Sub_Event Type: Results were Low, with Accuracy 30%, F1 score 0.02, though improved performance for some classes (Somewhat lower than, sub-event-type).

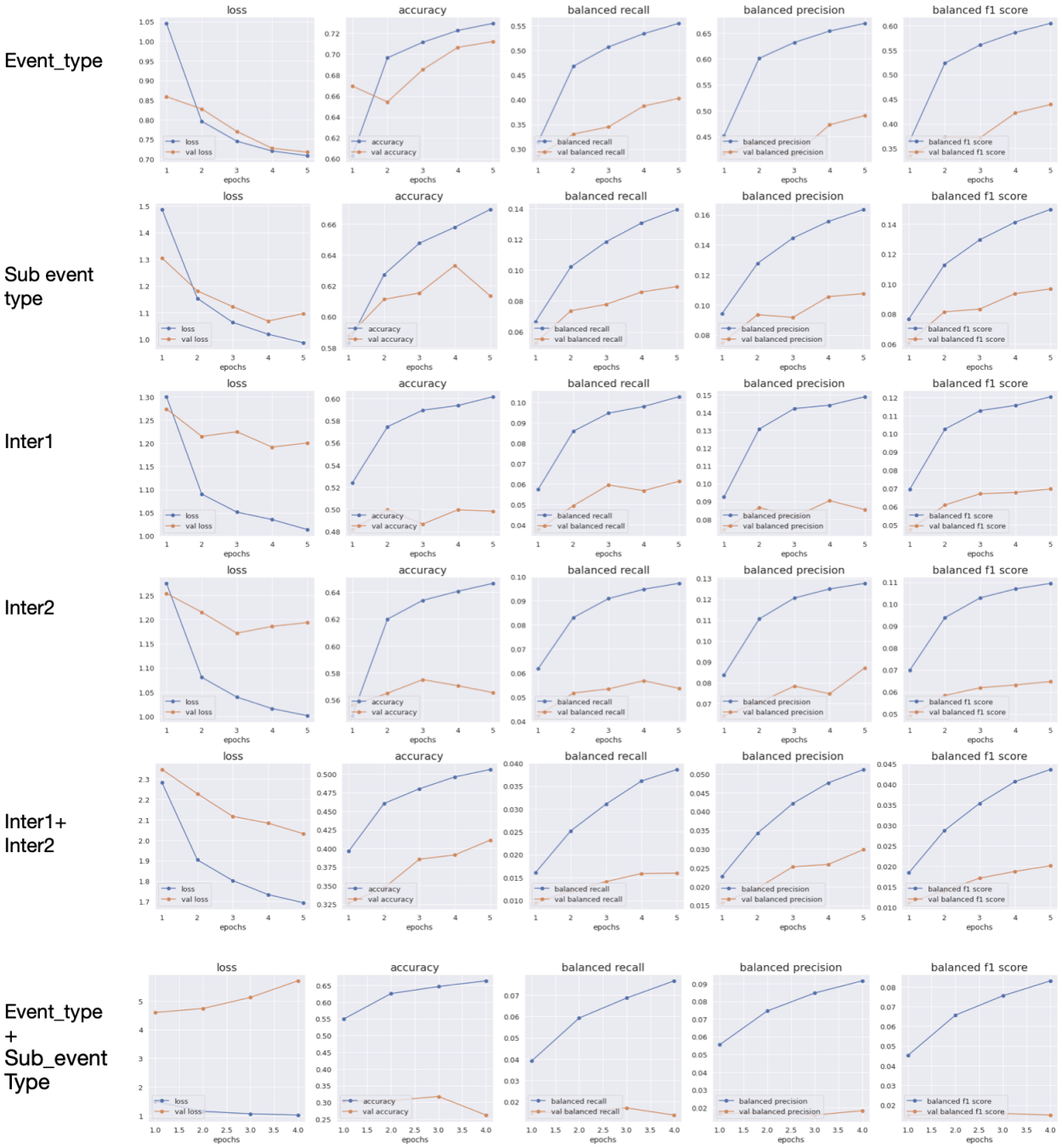


Figure 1: Results for single-col and combined-col multi-class classification

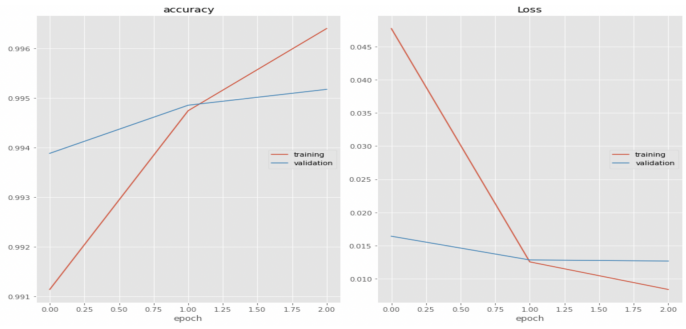


Figure 2: NER Training

For Fatalities: Regression Results Learning plateaued, and there was no reduction in loss. For Fatalities and Location, NER gave promising results as milestone 2 with 56% effective accuracy, and an F1 score of 0.34

Interaction is a derived field and hence not considered. Actor 1 and Actor 2 are only partially derived from Inter 1 and Inter 2 respectively.

5.2 Summarization Results

Bleu Score: 0.014

Rouge Score: Please refer 3

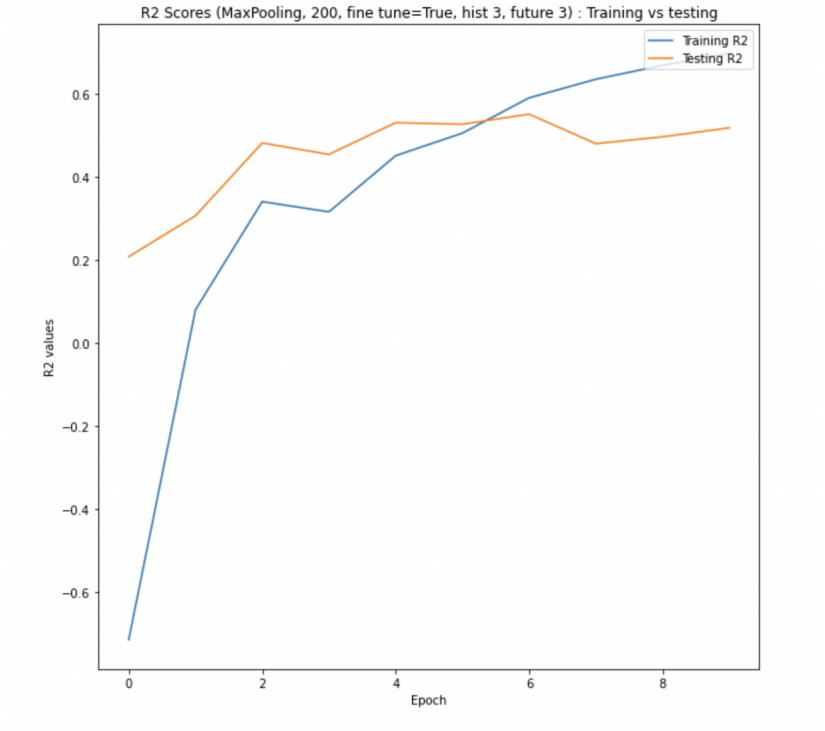
The bleu score has dropped from 0.24 to 0.014 as so have the rouge metrics. The drop is significant but the output generated from this better is better in human interpretable and understanding way. The outputs are much better phrased and more context keywords and features sets is captured. The context getting captured can be attributed to the self attention mechanism of the transformers.

Type	Recall	Precision	f1
Rouge-1	0.21	0.13	0.16
Rouge-2	0.04	0.03	0.03
Rouge-l	0.17	0.11	0.13

Table 2: Rouge metrics

5.3 Event Prediction

Results: Refer to Table 3 3



R squared scores for Hist 3, Lead time 3 with Fine Tuning and 10 epochs

5.3.1 Comparing results

The baseline system was a Bi LSTM with 4 layers. The problem was formulated as single step univaraite time series forecasting problem. With a windows size of 5 and lead time of 1, the R squared score of 0.2384 was achieved. With the same model, increasing windows size to 10 with same lead time resulted in R squared score of .3248.

History	Future	Fine Tune	Score
3	1	True	0.44
3	3	True	0.51
3	20	True	0.19
3	20	False	0.28
3	20	False	0.24
3	20	False	0.036

Table 3: R squared score

The scores with Transformer based model outperform the earlier results. With a window size of merely 3. Even with increased Lead time the R squared score are much better(Table 3: with Fine Tune = True). As earlier model just relied on date and no textual information and this model using textual information, the observed improvement is quite understandable. Another thing to note is that transformers have been pretrained on large corpus of data. So the model is inherently more sophisticated.

6 Discussion and Error Analysis

6.1 Information Extraction

- Lower results on multiclass classification, both single column (Inter1 f1:0.07, Inter2 f1:0.06) and dual column (Event_type and Sub_event type f1:0.02) can be attributed to less samples from result classes or increased number of effective classes.
- Dependencies of certain columns can be leveraged for prediction. Rather than consolidating Event_type and Sub_event type, as subeventtype has a reasonable accuracy of 0.61, it can be inferred first, and then used to map the parent class Event Type. Also, Inter1Inter2 consolidation is a good way to predict both classes, and can be used to infer Actors.
- In our current assumption, fatalities which are reflected in compound phrases like ”3 civilians and 2 soldiers” for 5 fatalities are not accounted for, and Location field values are sometimes not specific as required. These might need a better contextual approach than NER, or rather a more text-to-text prediction approach to do that.
- Actor 1, Actor 2 are partially derived from Inter1 and Inter2, but also have a textual value pertaining to the specific region or event. These have not been accounted for, and needs including the detected location while training, and more data from external sources (using the given date and source fields) to augment the given data, to build enough context. In addition, a more text-to-text prediction approach, possibly with attention based model might work better in this case as well.

6.2 Summary Generation

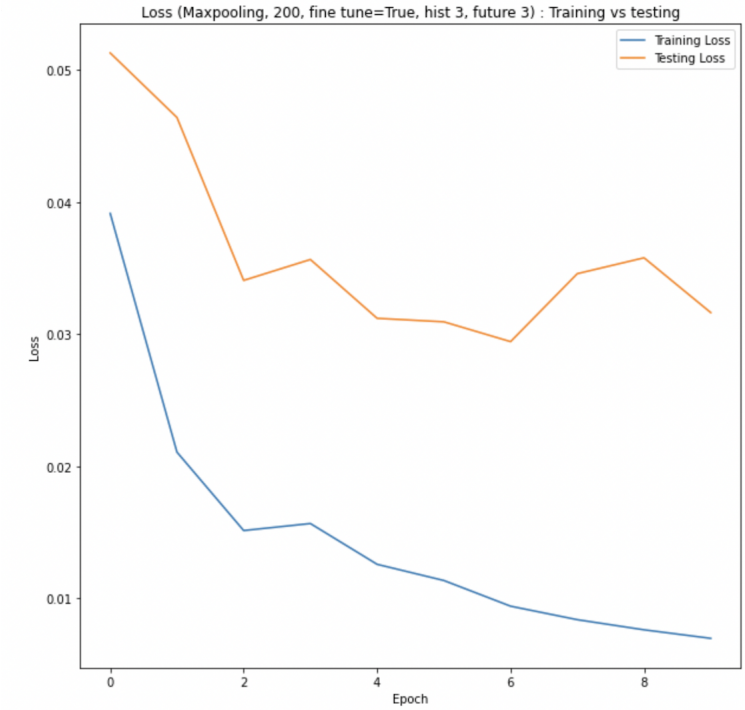
The model was inaccurate in identifying dates and properly mapping it. We can use a different tokenization mechanism to encode dates.

Since this model is more computional heavy compared to the previous milestone, the training size dataset used for the model was 10% of the available and also we limited training till 10 epochs. Using the entire dataset and more training on more epochs can increase the efficiency of the model.

6.3 Event Prediction

The current model uses only the text field, and is not incorporating the temporal nature of events. Additionally max pooling of embeddings gave descent results but is not the best way. Experiments can also be conducted with sum pooling and multi headed attention.

Additionally, the model uses all data irrespective of events taking place in different geography. More localized models can be developed. These 2 measures should improve the results and are a evident source of error in the current model.



Loss over 10 epoch for Hist 3, Lead time 3 with Fine Tuning

7 Conclusion

For Information Extraction, we weren't able to consolidate a single model to predict all target variables, hence given that all columns are different data-types, it has been promising to consider the best model for specific datatype separately. In this case, we came up with a multiclass approach for categorical columns and NER for attributed values like location, and were able to get reasonable accuracy, but below average F1 scores.

For Summarization, we came up with a model which was capturing small subset of feature in the output even with low number of epochs. Considering the quality of output we are getting even on a small subset, we can get better results in future when trained on entire dataset and with even more epochs.

For Event Prediction, we trained BERT transformer using pooled version of individual embeddings. The approach used to generate embeddings although gave good results but is by no means state of the art. The results have shown that in order to predict far in future we need large amount of historical data. But even then there is a limit on accuracy that can be achieved. However predicting in near future can be done with much more confidence.

8 Work Distribution

- Prateeksha Singh: Information Extraction, meetings
- Aditya Jagtap: Summarization
- Sanyam Gupta: Event Prediction

References

David Nadeau, Satoshi Sekine 2015. *A survey of named entity recognition and classification*, https://nlp.cs.nyu.edu/sekine/papers/li07.pdf

Jay Alammar 2014. *The Illustrated Word2Vec*, https://jalammar.github.io/illustrated-word2vec/

Rrubaa Panchendrarajan, Aravindh Amaresan 2018. *Bidirectional LSTM-CRF for Named Entity Recognition*, https://aclanthology.org/Y18-1061.pdf

Yoon Kim 2014. *Convolutional Neural Networks for Sentence Classification*, https://arxiv.org/pdf/1408.5882.pdf

Chris McCormick, Nick Ryan 2019. *BERT Fine-Tuning Tutorial with PyTorch*, https://mccormickml.com/2019/07/22/BERT-fine-tuning/

Varun Singh 2019. *Transformers for NER*, https://keras.io/examples/nlp/nertransformers/

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin 2017. *Attention Is All You Need*, https://arxiv.org/abs/1706.03762

Lu Meng and Rohini K. Srihari 2019. *Leveraging Heterogeneous Data Sources for Civil Unrest Prediction*, http://sbp-brims.org/2019/proceedings/papers/working$_papers/LuMeng_paper_1$13.*pdf*

Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang and Zhifang Sui 2017. *Table-to-text Generation by Structure-aware Seq2seq Learning*, https://arxiv.org/pdf/1711.09724v1.pdf

- Y. Hua, Z. Zhao, R. Li, X. Chen, Z. Liu and H. Zhang, "Deep Learning with Long Short-Term Memory for Time Series Prediction," in IEEE Communications Magazine, vol. 57, no. 6, pp. 114-119, June 2019, doi: 10.1109/MCOM.2019.1800155.
- Lu Meng and Rohini K. Srihari. Leveraging Heterogeneous Data Sources for Civil Unrest Prediction. 2019
- Lu Meng and Rohini K. Srihari. Increasing Lead Time and Granularity of Civil Unrest Prediction through Time Series Data 2019.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez and L. Kaiser. Attention is all you need. 2017.
- BERT: Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019