CS60075: Natural Language Processing

Group Project (Group: 30, Project: P6)

EVENT EXTRACTION

Dhruv Rathi(20EC10098)
Agnibha Sinha(20EC10001)
Faizan Ahmed(20MF10010)

Task Definition:

The task was to classify sentences based on the events involved in them. The approach was to implement a multi-class BERT classifier model for classifying the sentences into 25 unique classes thereby giving us information regarding the event being talked about in the sentence.

Description of the dataset:

The dataset consisted of multiple columns but our task focused mainly on 2 columns, the **notes** and the **sub_event_type**. The dataset contained **231821 rows**
The notes column consisted of sentences of **maximum length = 490**.
The sub_event_type column consisted of **25 unique labels** in the dataset but there were **30 labels in the test data.**
The labels not included in the dataset used for training but present in the test dataset were:

1) NATURAL_DISASTER
2) MAN_MADE_DISASTER
3) ATTRIB
4) DIPLO
5) ORG_CRIME

The **frequency of the 25 labels** in the given dataset are as follows:

```
Peaceful protest : 99445
Air/drone strike : 16080
Violent demonstration : 8714
Shelling/artillery/missile attack : 27168
Armed clash : 24657
Mob violence : 10872
Looting/property destruction : 4369
Attack : 10630
Disrupted weapons use : 2313
Arrests : 1749
Protest with intervention : 5941
Abduction/forced disappearance : 2151
Remote explosive/landmine/IED : 7487
Excessive force against protesters : 345
Government regains territory : 1785
Grenade : 824
Agreement : 236
Other : 1304
Change to group/activity : 4687
Non-violent transfer of territory : 188
Suicide bomb : 75
Headquarters or base established : 156
Sexual violence : 137
Non-state actor overtakes territory : 504
Chemical weapon : 4
```

Approach & Results:

The BERT model was used for the classification task. The model took the word embeddings as input along with the attention masks and had 10 hidden layers before finally returning the output layer which was a list of length 25 which is indicative of the probabilities of each label corresponding to the given word embedding.

The sentences were first tokenized using a bert tokenizer and then used for training and validation. The **bert tokenizer** performs the following tasks:

1) Makes the length of the sentences equal to 490(max_length) by padding.
2) Makes the first token CLS which contains the word embeddings and the last token SEP.
3) Uses attention masks to pass on the information on which tokens are relevant.

The labels were encoded as follows:

```
'Abduction/forced disappearance': 11,
 'Agreement': 16,
 'Air/drone strike': 1,
 'Armed clash': 4,
 'Arrests': 9,
 'Attack': 7,
 'Change to group/activity': 18,
 'Chemical weapon': 24,
 'Disrupted weapons use': 8,
 'Excessive force against protesters': 13,
 'Government regains territory': 14,
 'Grenade': 15,
 'Headquarters or base established': 21,
 'Looting/property destruction': 6,
 'Mob violence': 5,
 'Non-state actor overtakes territory': 23,
 'Non-violent transfer of territory': 19,
 'Other': 17,
 'Peaceful protest': 0,
 'Protest with intervention': 10,
 'Remote explosive/landmine/IED': 12,
 'Sexual violence': 22,
 'Shelling/artillery/missile attack': 3,
 'Suicide bomb': 20,
 'Violent demonstration': 2
```

The dataset was split into training and validation sets. The **training dataset contained 52%** of the data. The size of the training dataset was kept small due to restrictions on computational power.

<u>Training:</u>
The **word embeddings** of the sentences along with the **label encodings** and the **attention masks** were passed to the pretrained BERT model for training.
The model calculated a **training loss of 0.360354**
Due to the restrictions on the time for which GPU was available, we ran just 1 epoch. Also the batch size was kept as 3 due to limitations on memory available.

<u>Validation:</u>
After the training was completed, the model was validated using the validation dataset.

The validation loss was 0.200955
The weighted F1 score was 0.956766

The accuracy obtained for the different classes are listed below:

```
Class: Peaceful protest
Accuracy: 39522/39778
Class: Air/drone strike
Accuracy: 6376/6432
Class: Violent demonstration
Accuracy: 3137/3485
Class: Shelling/artillery/missile attack
Accuracy: 10702/10867
Class: Armed clash
Accuracy: 9145/9863
Class: Mob violence
Accuracy: 3946/4349
Class: Looting/property destruction
Accuracy: 1587/1747
Class: Attack
Accuracy: 3934/4252
Class: Disrupted weapons use
Accuracy: 861/925
Class: Arrests
Accuracy: 513/700
Class: Protest with intervention
Accuracy: 2254/2376
Class: Abduction/forced disappearance
Accuracy: 810/860
Class: Remote explosive/landmine/IED
Accuracy: 2940/2995
Class: Excessive force against protesters
Accuracy: 0/138
Class: Government regains territory
Accuracy: 677/714
```

```
Class: Grenade
Accuracy: 230/330
Class: Agreement
Accuracy: 45/94
Class: Other
Accuracy: 439/522
Class: Change to group/activity
Accuracy: 1840/1875
Class: Non-violent transfer of territory
Accuracy: 0/75
Class: Suicide bomb
Accuracy: 0/30
Class: Headquarters or base established
Accuracy: 0/62
Class: Sexual violence
Accuracy: 0/55
Class: Non-state actor overtakes territory
Accuracy: 0/202
Class: Chemical weapon
Accuracy: 0/2
```

Testing:

The model was then saved and later used on a test dataset consisting of 1023 unlabeled sentences. The predicted labels were obtained from the model and later we received the correct labels and compared it to the predicted labels.

It was found that the **446 labels were correctly predicted.**

The test dataset had 5 extra labels which were not included during training. Also we found an ambiguity in the test dataset where sentences with ids 142, 205, 737 and 957 were missing from the test file with labels. Also the ids were jumbled in between and not sequential.

Also the other labels were similar but not identical to the ones in the training. So we had to map them through the same encodings manually.

The label encodings consistent with the training labels are attached below:

```
label_test_dict =
{'AIR_STRIKE':1,'NATURAL_DISASTER':25,'FORCE_AGAINST_PROTEST':13,'NON_STAT
E_ACTOR_OVERTAKES_TER':23,'AGREEMENT':16,'CHEM_WEAP':24,
'PEACE_PROTEST':0,'GOV_REGAINS_TERIT':14,'DISR_WEAP':8,'PROPERTY_DISTRUCT'
:6,'OTHER':17,'CHANGE_TO_GROUP_ACT':18,'GRENADE':15,'VIOL_DEMONSTR':2,
'MAN_MADE_DISASTER':26,'ATTRIB':27,'MOB_VIOL':5,'ATTACK':7,'ARMED_CLASH':4
,'ART_MISS_ATTACK':3,'NON_VIOL_TERRIT_TRANSFER':19,
'PROTEST_WITH_INTER':10, 'DIPLO':28, 'SUIC_BOMB':20,
'ARREST':9,'REM_EXPLOS':12, 'SEX_VIOL':22,
'ORG_CRIME':29,'ABDUCT_DISSAP':11, 'HQ_ESTABLISHED':21}
```

The labels not included in the dataset used for training but present in the test dataset were: NATURAL_DISASTER, MAN_MADE_DISASTER, ATTRIB, DIPLO, ORG_CRIME.

The accuracy of each class in the test data was found as follows:

```
PEACE_PROTEST : 53 / 61 = 0.8688524590163934
AIR_STRIKE : 31 / 36 = 0.8611111111111112
VIOL_DEMONSTR : 26 / 53 = 0.49056603773584906
ART_MISS_ATTACK : 27 / 36 = 0.75
ARMED_CLASH : 60 / 66 = 0.9090909090909091
MOB_VIOL : 8 / 17 = 0.47058823529411764
PROPERTY_DISTRUCT : 4 / 21 = 0.19047619047619047
ATTACK : 21 / 27 = 0.7777777777777778
DISR_WEAP : 45 / 58 = 0.7758620689655172
ARREST : 12 / 34 = 0.35294117647058826
PROTEST_WITH_INTER : 19 / 22 = 0.8636363636363636
ABDUCT_DISSAP : 16 / 20 = 0.8
REM_EXPLOS : 35 / 36 = 0.9722222222222222
FORCE_AGAINST_PROTEST : 0 / 23 = 0.0
GOV_REGAINS_TERIT : 28 / 38 = 0.7368421052631579
GRENADE : 32 / 48 = 0.6666666666666666
AGREEMENT : 0 / 31 = 0.0
OTHER : 2 / 8 = 0.25
CHANGE_TO_GROUP_ACT : 27 / 30 = 0.9
NON_VIOL_TERRIT_TRANSFER : 0 / 21 = 0.0
SUIC_BOMB : 0 / 41 = 0.0
HQ_ESTABLISHED : 0 / 22 = 0.0
SEX_VIOL : 0 / 23 = 0.0
NON_STATE_ACTOR_OVERTAKES_TER : 0 / 24 = 0.0
CHEM_WEAP : 0 / 37 = 0.0
NATURAL_DISASTER : 0 / 37 = 0.0
MAN_MADE_DISASTER : 0 / 52 = 0.0
ATTRIB : 0 / 28 = 0.0
DIPLO : 0 / 44 = 0.0
ORG_CRIME : 0 / 29 = 0.0
```

Observations:

The model on validation performed really well on certain classes like “peaceful protest” whereas it did not fit to a few classes which had a lower frequency in the training dataset like “chemical weapon”. This was due to underfitting as the data for such classes was really low.

The model did not perform very well on the test dataset which may be due to the following:

1) The number of epochs was kept as 1 and the training data was also less due to limitations on available computational power. Thus the model did not fit very well.
2) The test data sets had some ambiguities as pointed out in the previous section.
3) The extra labels in the test data set were not possible to predict using this model.

Improvements:

We had initially tried to implement an LSTM classifier but found the BERT model more suitable. We further came to know about the ROBERTA, ALBERT models which would be even better for the task but did not get time to implement it.
The problem of classifying the labels not present in the training set could be solved using few shot or zeroshot classification.

Conclusion

Through this project we were able to successfully implement the BERT model and gain information about the type of event being talked about in various sentences. We learnt a lot during this project both implementation of concepts learnt in class and newer concepts. We would like to thank Prof. Sudeshna Sarkar for teaching us concepts in class which helped us in this project. We would also like to thank Alapan Kuila sir for constantly guiding us in this project.

Contributions

Dhruv Rathi - Training and Validation
Agnibha Sinha - Testing and Documentation
Faizan Ahmed - Preprocessing and Documentation

Links to Datasets

1) Original Dataset
2) Test Dataset with labels
3) Saved Model
4) Code for training
5) Code for testing