"MULTILINGUAL CHARACTERIZATION AND EXTRACTION OF NARRATIVES FROM ONLINE NEWS"

By: Muhammad Khubaib Mukaddam (mk07218),
Muhammad Shoaib Khursheed (mk07149),
Muminah Khurram (mk07521)

Problem Motivation & Challenges

Proliferation of Deceptive Content:

- Internet has become a conduit for both information sharing and the spread of deceptive content.
- Users exposed to manipulation and misinformation during major crises.

Challenges in Analysis:

- Analysts struggle to keep up with the volume and complexity of online news.
- Need for advanced tools to identify and characterize manipulation attempts effectively.

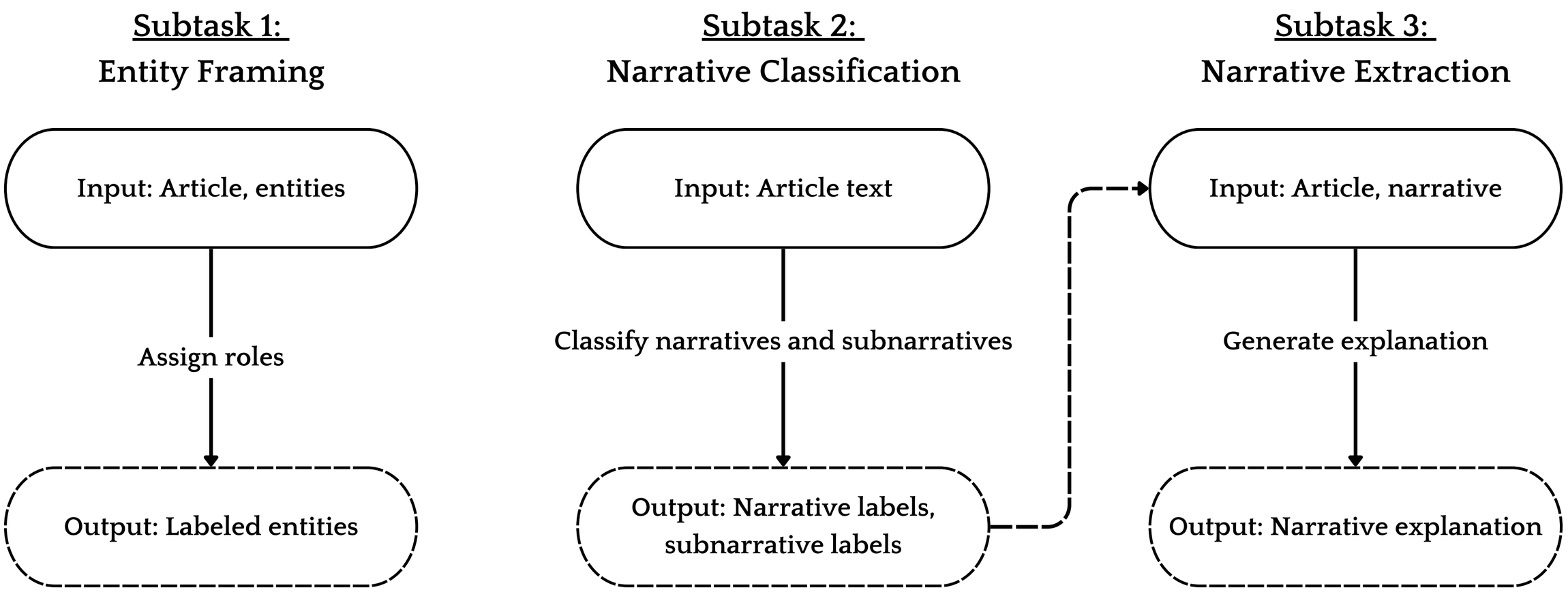
Topics:

- Climate Change
- Russia-Ukraine War

Languages:

- Hindi
- English
- Bulgarian
- Portugese

Task Division and Description



Example

Met Office Should Put 2.5°C 'Uncertainties' Warning on All Future Temperature Claims

It is "abundantly clear" that the Met Office cannot scientifically claim to know the cur- rent average temperature of the U.K. to a hundredth of a degree centigrade, given that it is using data that has a margin of error of up to 2.5°C, notes the climate journalist Paul Homewood. His comments follow recent disclosures in the Daily Sceptic that nearly eight out of ten of the Met's 380 measuring stations come with official 'uncertainties' of between 2-5°C. In addition, given the poor siting of the stations now and possibly in the past, the Met Office has no means of knowing whether it is comparing like with like when it publishes temperature trends going back to 1884.

There are five classes of measuring stations identified by the World Meteorological Office (WMO). Classes 4 and 5 come with uncertainties of 2°C and 5°C respectively and account for an astonishing 77% of the Met Office station total. Class 3 has an uncertainty rating of 1°C and accounts for another 8.4% of the total. The Class ratings identify potential corruptions in recordings caused by both human and natural involvement. Homewood calculates that the average uncertainty across the entire database is 2.5°C. In the graph below, he then calculates the range of annual U.K. temperatures going back to 2010 incorporating the margins of error.
...

Subtask 1: Entity Framing

- Met Office: Antagonist-[Deceiver]
- Paul Homewood: Protagonist-[Guardian]
- Daily Sceptic: Protagonist-[Guardian]
- Christopher Booker: Protagonist-[Guardian,Virtuous]

Example

Met Office Should Put 2.5°C 'Uncertainties' Warning on All Future Temperature Claims

It is "abundantly clear" that the Met Office cannot scientifically claim to know the cur- rent average temperature of the U.K. to a hundredth of a degree centigrade, given that it is using data that has a margin of error of up to 2.5°C, notes the climate journalist Paul Homewood. His comments follow recent disclosures in the Daily Sceptic that nearly eight out of ten of the Met's 380 measuring stations come with official 'uncertainties' of between 2-5°C. In addition, given the poor siting of the stations now and possibly in the past, the Met Office has no means of knowing whether it is comparing like with like when it publishes temperature trends going back to 1884.

There are five classes of measuring stations identified by the World Meteorological Office (WMO). Classes 4 and 5 come with uncertainties of 2°C and 5°C respectively and account for an astonishing 77% of the Met Office station total. Class 3 has an uncertainty rating of 1°C and accounts for another 8.4% of the total. The Class ratings identify potential corruptions in recordings caused by both human and natural involvement. Homewood calculates that the average uncertainty across the entire database is 2.5°C. In the graph below, he then calculates the range of annual U.K. temperatures going back to 2010 incorporating the margins of error.
...

Subtask 2: Narrative Classification

- NARRATIVE: Questioning the measurements and science.
- SUB-NARRATIVE: Methodologies/metrics used are unreliable/fault.

Example

Met Office Should Put 2.5°C 'Uncertainties' Warning on All Future Temperature Claims

It is "abundantly clear" that the Met Office cannot scientifically claim to know the cur- rent average temperature of the U.K. to a hundredth of a degree centigrade, given that it is using data that has a margin of error of up to 2.5°C, notes the climate journalist Paul Homewood. His comments follow recent disclosures in the Daily Sceptic that nearly eight out of ten of the Met's 380 measuring stations come with official 'uncertainties' of between 2-5°C. In addition, given the poor siting of the stations now and possibly in the past, the Met Office has no means of knowing whether it is comparing like with like when it publishes temperature trends going back to 1884.

There are five classes of measuring stations identified by the World Meteorological Office (WMO). Classes 4 and 5 come with uncertainties of 2°C and 5°C respectively and account for an astonishing 77% of the Met Office station total. Class 3 has an uncertainty rating of 1°C and accounts for another 8.4% of the total. The Class ratings identify potential corruptions in recordings caused by both human and natural involvement. Homewood calculates that the average uncertainty across the entire database is 2.5°C. In the graph below, he then calculates the range of annual U.K. temperatures going back to 2010 incorporating the margins of error.
...

Subtask 3: Narrative Extraction

Paul Homewood claims that the Met Office is misleading the public about current UK temperatures by not disclosing a margin of error of up to 2.5° C. The Daily Sceptic reports that most of the Met Office's 380 stations provide inaccurate measurements. Additionally, Christopher Booker argues that official reports have been repeatedly falsified to indicate climate warming. Thus, the Met Office cannot conclude with scientific certainty that the climate is becoming warm

Data Augmentation - Subtask 1 & 2

The initial dataset provided for the tasks was relatively small, which posed a significant challenge for training robust and generalizable models.
For Subtask 1 and 2 we audemnted the data using Gemini Developer API

Techniques Used:

- Subtask 1: For every entry paraphase the data and make duplicate entry for each label.
- Subtask 2: For every langauge dataset translated data o Hindi and reverted back to English.

Sampling for Subtask 1

Considering there was a data imbalance, conducted literature review to understand techniques to understand how to tackle this. Tried both oversampling and undersampling .

Model and Experiments - Subtask 1

Observations:

- BART (CNN) achieved the best performance among tested models.
- BART Large also performed good
- BERT-base-uncased underperformed severely.
- DistilBERT-base-uncased performed a little better than Bert yet still was pretty low
- Even after using **Contrastive Loss,** there was no prominent effect on our results.

Model	F1
DistilBERT-base-uncased	**0.13190**
BERT-base-uncased	**0.12090**
BART-CNN ★	**0.24180**
BART-Large	**0.21980**

Model and Experiments - Subtask 1

Result

Rank	Team	**Exact Match Ratio**	micro P	micro R	micro F1	Accuracy for main role
1	nihao	**0.45050**	0.50550	0.46000	0.48170	0.87910
2	QUST	**0.38460**	0.40660	0.37000	0.38740	0.80220
3	DEMON	**0.35160**	0.41760	0.38000	0.39790	0.86810
4	mbzuaijcbcruz	**0.30770**	0.13040	0.12000	0.12500	0.58240
5	NarrativeMiners	**0.24180**	0.28570	0.26000	0.27230	0.80220
6	news88readers	**0.21980**	0.27470	0.25000	0.26180	0.78020
7	Baseline	**0.12090**	0.12090	0.11000	0.11520	0.80220
7	Dhananjaya	**0.12090**	0.12090	0.11000	0.11520	0.80220
7	PAI	**0.12090**	0.12090	0.11000	0.11520	0.80220
7	gowithnlp	**0.12090**	0.12090	0.11000	0.11520	0.80220
8	Tuebingen	**0.09890**	0.14410	0.16000	0.15170	0.74730
9	KevinMBZUAI	**0.05490**	0.27370	0.26000	0.26670	0.26370
9	MikasaAckerman	**0.05490**	0.06990	0.10000	0.08230	0.80220

Model and Experiments - Subtask 2

Version 1

Independent:

No Heirchy:

1 model for both Narratives and Subnarratives

Model	F1
BERT	0.008

Model and Experiments - Subtask 2

Version 2

Heirchal: Narratives, Then Subnarratives

1 model for both Narratives and Subnarratives

Model	Percesion	Recall	F1
BERT	0.12	0.2	0.09

Model and Experiments - Subtask 2

Version 3

Three step heirchy

Heirchal: Groups, Narratives, Then Subnarratives

1 model for both Narratives and Subnarratives

Model	Percesion	Recall	F1
BERT	0.16	0.4	**0.12**

Model and Experiments - Subtask 2

Version 4

Three step heirchy

Heirchal: Groups, Narratives, Then Subnarratives

5 model: 1 to predict group, 2 for Narratives and 2 for Subnarratives

Model	Percesion	Recall	F1
BERT	0.2	0.45	0.171

Model and Experiments - Subtask 2

Result

Taking time to update leaderboard

Dear team 'NarrativeMiners', thanks for your submission!

The file 9.txt has been uploaded.

Scoring your file...

```
2024-11-29 09:43:27,592 - INFO - Evaluation Results:
F1@coarse: 0.215 (0.374)
F1@fine: 0.171 (0.376)
```

Date	F1 macro coarse	F1 st. dev. coarse	**F1 macro fine**	F1 st. dev. fine
November 29 9:43:26	0.21500	0.37400	**0.17100**	0.37600

Experimental Setup - Subtask 3

Data:

- Used the provided training set of news articles with gold standard explanations.

Training:

- Fine-tuned each model on the task-specific data.
- Employed standard hyperparameters unless otherwise specified.

Evaluation:

- Metrics: Precision, Recall, F1 Score (Macro), using BertScore for similarity assessment between generated and gold explanations.

Model and Experiments - Subtask 3

Observations:

- BART (CNN) achieved the best performance among tested models; effectively captures contextual information for text generation.
- BART CNN and BART Large have a small tradeoff in score but logically different.
- GPT and Flan underperformed compared to BART; struggle with generating coherent explanations.
- LLama (all versions, all models) encountered CUDA out-of-memory errors; experiments still are ongoing.

Model	Percesion	Recall	F1 Score (Macro)
BART-CNN★	0.7286	0.7488	0.7385
BART Large	0.76180	0.69615	0.72723
GPT-2	0.5854	0.6964	0.6360
Flan-T5	0.6727	0.6217	0.6456
LLaMA 3.2 1b	-	-	-

Results - Subtask 3

Final Submission:

- Model: Bart CNN Large
- Specifications: 7 epochs, 4 batch size
- Trained on: BertScore - F1 (macro)
- Dataset Split: 77 train, 12 Validation

English - Subtask 3

Rank	Team	Precision	Recall	F1 macro
1	Genadium	0.75261	0.73397	**0.74304**
2 ★	NarrativeMiners	0.72860	0.74883	**0.73848**
3	insun606	0.72366	0.70875	**0.71595**
4	DUTtask10	0.70203	0.69695	**0.69936**
5	telorbulat	0.66421	0.70441	**0.68306**
6	Baseline	0.65540	0.67957	**0.66719**
6	news88readers	0.65540	0.67957	**0.66719**

Approaches & Litrature Review - Subtask 3

Approach 1: CLEF 2024 SimpleText Track

LLama (2 & 3) and GPT 3.5 Turbo

Approach 2: Team Sharingans at SimpleText

GPT 3.5 Turbo

Successes:

- Demonstrated that fine-tuned transformer models can effectively perform narrative extraction.
- BART shows promise as a leading model for this task.

Next Steps:

- Resolve hardware issues to experiment with larger models (e.g., LLaMA).
- Explore reinforcement learning techniques to further align generated content with source text.
- Investigate additional data augmentation methods and cross-lingual training to enhance model robustness.

THANK YOU FOR LISTENING!