CLEARumor at SemEval-2019 Task 7: ConvoLving ELMo Against Rumors

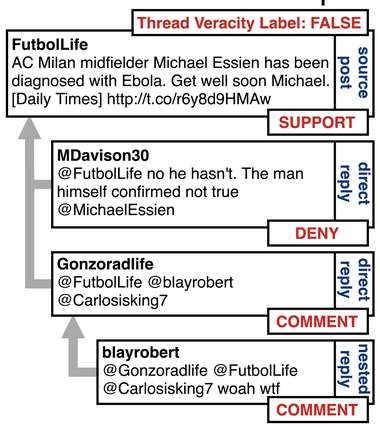
Summary

- **Second place** in the RumourEval 2019 competition
- Pre-trained **ELMo embeddings** [1] to integrate general language understanding
- **CNN-based** model with auxiliary input features

RumourEval 2019

- Labelled collection of comment threads from Twitter and Reddit
- Subtask A (SDQC): classify each comment as **support**, **deny**, **query**, or **comment** towards the rumor in its thread's source post
- Subtask B (Veracity): classify the rumor expressed in the thread's source post as **true**, **false**, or **unverified**

Example



Dataset

Subtask A		S	D	Q	C	Σ
Train	Twitter	910	344	358	2907	**5217**
	Reddit	15	34	37	612	
Dev	Twitter	94	71	106	778	**1485**
	Reddit	8	11	14	403	
Test	Twitter	141	92	62	771	**1827**
	Reddit	16	9	31	705	
Σ		**1184**	**561**	**608**	**6176**	**8529**

Subtask B		T	F	U	Σ
Train	Twitter	137	62	98	**327**
	Reddit	7	17	6	
Dev	Twitter	8	12	8	**38**
	Reddit	2	7	1	
Test	Twitter	22	30	4	**81**
	Reddit	9	10	6	
Σ		**185**	**138**	**133**	**456**

Model

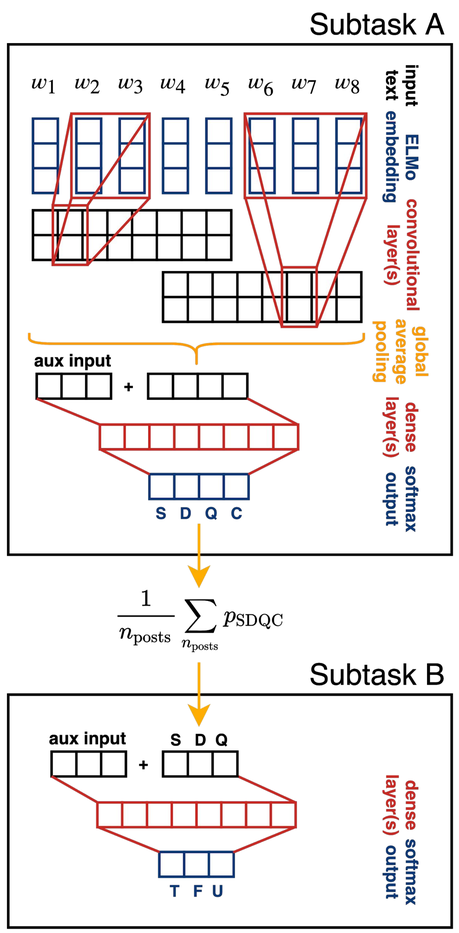
- ReLU activation
- 1D-convolution (kernel sizes 2 & 3)
- Batch norm
- L_2-regularization
- Adam for training

Preprocessing

- Lowercase everything
- Remove user handles & URLs
- Transform hash tags into words
- Limit character repetitions to 3
- Truncate after 32 tokens

Auxiliary Input

Subtask	A	B
On Twitter or Reddit?	✓	✓
User verified?	✓	✓
Number of followers	✓	✓
Number of followees	✓	✓
Ratio followers/followees	✓	✓
Similarity to source post	✓	
Source, reply, nested post?	✓	
Media attached?		✓
Upvote-to-downvote ratio		✓
Fraction of direct replies		✓
Fraction of nested replies		✓



Evaluation

Subtask A	Dev	Test					CV
	Macro-F_1	Macro-F_1	S-F_1	D-F_1	Q-F_1	C-F_1	Macro-F_1
Always Comment	22.1	22.3	0.0	0.0	0.0	89.4	—
Submitted	41.3	37.4	**46.7**	0.0	11.7	**91.2**	—
$\text{CLEAR}^{\text{aux}}$	**44.8**±0.6	42.7±0.6	29.6±0.6	**17.8**±2.4	43.9±1.0	79.5±1.3	47.1±4.5
$\text{CLEAR}^{\text{aux}}_{\text{MLP}}$	42.2±1.2	40.7±1.6	30.7±2.7	0.0±0.0	**51.6**±3.2	80.5±2.7	44.7±4.2
$\text{CLEAR}_{\text{CNN+MLP}}$	39.7±2.0	39.0±2.2	16.2±2.3	14.8±3.4	41.0±6.7	84.0±2.6	43.3±4.5
$\text{CLEAR}^{\text{aux}}_{\text{CNN+MLP}}$	42.9±2.2	**44.6**±2.6	34.6±3.7	15.4±3.1	42.2±8.3	86.1±1.1	**47.2**±3.8

Results averaged over 10 runs. CV is 10-fold cross validation.

- "Always Comment": **baseline** always predicting the most common class
- "Submitted": **preliminary results** we submitted to RumourEval 2019
- $\text{CLEAR}^{\text{aux}}$: ELMo embeddings + **auxiliary** input with **linear** projection
- $\text{CLEAR}^{\text{aux}}_{\text{MLP}}$: ELMo embeddings + **auxiliary** input with **dense** layers
- $\text{CLEAR}_{\text{CNN+MLP}}$: ELMo embeddings with **convolutional** and **dense** layers
- $\text{CLEAR}^{\text{aux}}_{\text{CNN+MLP}}$: ELMo embs + **aux** input with **convolutional** and **dense** layers

Subtask B	Dev		Test		CV	
	Macro-F_1	RMSE	Macro-F_1	RMSE	Macro-F_1	RMSE
Submitted	41.7	0.743	28.6	0.764	—	—
$\text{CLEAR}_{\text{Subtask-B}}$	35.4±0.5	**0.676**±0.005	**30.1**±0.8	**0.754**±0.005	**26.7**±13.4	**0.733**±0.113
$\text{CLEAR}_{\text{NileTMRG}}$	**53.5**	0.761	18.6	0.846	—	—

Results averaged over 10 runs. CV is 10-fold cross validation.

- "Submitted": **preliminary results** we submitted to RumourEval 2019
- $\text{CLEAR}_{\text{Subtask-B}}$: **our Subtask-B** system on our subtask-A predictions
- $\text{CLEAR}_{\text{NileTMRG}}$: the **NileTMRG** [2] system on our subtask-A predictions

References

[1] Peters et al (2018). Deep Contextualized Word Representations. NAACL-HLT.

[2] Enayet & El-Beltagy (2017). NileTMRG at SemEval-2017 Task 8: Determining Rumour and Veracity Support for Rumours on Twitter. SemEval@ACL.

İpek Barış ibaris@uni-koblenz.de

Lukas Schmelzeisen lukas@uni-koblenz.de

Steffen Staab staab@uni-koblenz.de

Institute for Web Science and Technologies (WeST), University of Koblenz-Landau, Germany

Source code at: `https://github.com/Institute-Web-Science-and-Technologies/CLEARumor`