Emotion Recognition in Online Text Chats

- Recognizing emotions is an ambiguous process with high dependence on the contextual information both on the utterance and dialog level.
- State-of-the-art approaches exploit whole dialog knowledge and have difficulties capturing temporal dependencies over long horizons.
- Accurate real-time emotion recognition is paramount importance for early identification of cyberbullying and suicidal ideation in Online Social Networks (OSNs).

Dataset

- We utilize a well-known benchmark for emotion recognition, namely the IEMOCAP dataset.
- The dataset has been collected by emulating conversations in a controlled environment in order to study expressive human behaviors.
- The conversations have been performed by ten unique speakers over five dyadic sessions in both a scripted but also an improvisation manner with various audio-visual modalities being recorded.
- Each utterance in the dataset is labeled by three human annotators using categorical labels; these include angry, sad, happy, frustrated, excited, neutral as well as other categories which we omit in this study.
- The available annotation has been performed by three annotators who assess the emotional states of the speakers taking into consideration dialog context.
- We only utilize the textual modality (prevalent form in OSNs) and the label information derived by performing majority voting.
- The dataset contains 151 conversations with a total number of 10,039 utterances. However, only 7,380 utterances contain the six types of emotions we retain in this study.

Self-Attentive Emotion Recognition Network

For the first time in the literature, we introduce a self-attentive hierarchical encoder network that is capable of extracting salient information on both the individual utterance level as well as the level of the dialog context, as it has evolved until any given time point.

- A trainable Word2Vec embeddings mechanism is presented with the input sequence.
- A bidirectional Long Short-Term Memory (BiLSTM) is used to capture the salient linguistic information contained within each utterance.
- A Gated Recurrent Unit (GRU) that performs dialog context-level representation to allow for capturing the salient dynamics over the whole dialog span.
- A self-attention layer on top of the dialog-level GRU network.

- The devised model is trained in an end-to-end fashion.
- The employed training objective function is the categorical cross-entropy of the model.
- The employed training objective function is the categorical cross-entropy of the model.

Experimental Evaluation

| SVM  | 0.315 (±0.00) | 0.484 (±0.00) | 0.325 (±0.00) | 0.316 (±0.00) |
| BiLSTM | 0.477 (±0.01) | 0.471 (±0.02) | 0.459 (±0.01) | 0.465 (±0.01) |
| BiLSTM$_{M_0}$ | 0.516 (±0.02) | 0.516 (±0.02) | 0.501 (±0.02) | 0.509 (±0.02) |
| SERN | 0.522 (±0.02) | 0.544 (±0.02) | 0.517 (±0.02) | 0.530 (±0.02) |

Acknowledgments

This research is supported by the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie ENCASE project (Grant Agreement No. 691025).