**Selection of pre-trained models for dimensional emotion theory parameter regression tasks**

The purpose of this section is to narrow the scope of potential neural network architectures for assessment for a regression task to generate dimensional emotional theory parameters. This task has been broken down into two phases:

*Phase 1:* Identification of architectures commonly used in relevant literature and critical evaluation of their features with regards to the specific task.

*Phase 2:* Using the most suitable identified architecture from phase 1, assess several pre-trained models which fall into this category for the regression task.

*Evaluation of common architectures in literature for sentiment regression task*

Sentiment is fundamentally linked to context and thus models which successfully capture sentiment are likely to contain characteristics which enable the capture of long-term dependencies. Neural network architectures are widely utilised for such purposes. Their suitability is largely attributable to aspects of their architecture such as their:

*Ability to handle sequential data*: As language is sequential in nature, sentiment is often linked to word order. As neural networks handle data sequentially, their outcomes are often improved compared to more simplistic models.

*Robustness against noise:* Language is inherently noisy and variable due to variance in vernacular across a population in addition to relatively frequent errors in spelling or grammar. Such an effect is known to be more prominent in online content, where such features are sometimes used to add nuance to pragmatics. In the context of a problem set where there is potential for such features to add insight, this feature of such architectures may possess limitations if this information cannot be extracted in another manner.

Several models which would fall under this category are prominently found throughout literature evaluating sentiment analysis methodologies and specifically in the domain of sarcasm detection due to these features. The three primary models identified were as follows:

LSTM: The LSTM mechanism of selective memory is made up of a cell state, hidden state, and gates. The NN models sequential data by propagating over time through the connection of sequential events using the hidden state. This component captures dependencies by considering both the previous step and current output:

However, a feature of such a mechanism result in all previous steps being considered in the current step when implemented in isolation due to the chain rule:

Because:

With the ultimate result being either vanishing or exploding gradients and thus limitations on the abilities to capture long term dependencies in isolation. The cell state mitigates this effect through the filtering of less relevant information from the memory through the forget gate. This information, in addition to the weights from the input gate enable the model to learn which time steps contain important information, resulting in weights for each time step being represented in proportion to their understood importance to the model. Literature evaluating this architecture is not consistent with regards to its assessment of the efficacy of such a mechanism; with some studies citing the model as effective to capture long-term dependencies, and others postulating that the mechanism may ‘dilute’ important information over time. Similar contrasting observations are found in the sentiment across literature with regards to their robustness against noise, with the former asserting that noise information is filtered efficiently during training and the latter arguing the converse and observing amplification of noise. In this context, what is traditionally regarded as noise may provide pragmatic cues with respect to potential sarcasm as discussed above, these cues may be indicative that context of the problem set plays a part in the efficacy of the architecture to model valid patterns in the data, which explains contrasting observations within literature. No works could be identified which assessed this hypothesis, however this may be an area for future research.

GRU: GRU models address vanishing and exploding gradients using a more simplistic mechanism than LSTM models. These models omit the cell state and regulate memory using gates. These architectures utilise an update gate to dictate the information which is retained from the previous step in series with a reset gate which dictates the information which should be eliminated. Compared to LSTMs for natural language processing tasks, due to their more simplistic architecture GRU models seem to perform better where shorter data sequences are used, possibly providing greater potential in the context of the short form content as is used for the problem set in question.

Given the lack of output gate, there is no internal memory present in these models, it would be expected that there would be some compromise in performance

State of the art models for sentiment analysis problem sets largely consist of iterations of transformer-based architectures.

**Optimisation Process for Sentiment-Aware Vector Space Modification Model**

**Architectures of Considered Emoji Sentiment Prediction Models**

**Survey Results**