How Does That Make You Feel? Understanding Abuse in Context by Studying Emotions and Sentiments in Tweets and Replies

Social networking platforms have suffered from a rise in collective behaviors such as trolling, bullying, hate speech, abusive language, and even threats and personal attacks. Understanding when a particular message can lead to a strong emotional response in other users can provide insight into the social, linguistic, and contextual factors which drive such interactions. In this work, we study the affects of tweet-reply pairs to evaluate the factors which can drive the polarity of responses and potentially be indicative of signs of abuse in social interactions. In contrast to previous work, our approach focuses on understanding sentiment and emotions of microblog content in the context of ongoing conversation. Additionally, our work expands on previous efforts by identifying factors which increase the prediction difficulty in order to identify anomalous social interactions which deviate from conversational norms.

We first present methods to predict if a tweet will elicit a positive or negative sentiment in its replies, as well as what specific emotion (e.g., fear, anger, disgust, etc.) it will elicit. For both the sentiment and emotion prediction tasks, we aim to predict both the presence of a particular evoked affect (classification) and the degree of response evoked (regression). To predict affects we employ pretrained GloVe embeddings for content representation and examine a variety of additional features such as user communication behavior, stylistic and psycholinguistic markers. We compare the performance of AdaBoost classification and regression models with LSTM-based deep learning methods. We leverage several sample weighting techniques to mitigate noise in the assignment of sentiment and emotion labels.

For the classification tasks, we achieve an F1 score of 0.78 for distinguishing tweets which will receive negative versus positive replies and F1 scores ranging from 0.58 to 0.68 for classifying the presence of each of the six emotion categories in the replies. We compare the performance of the classification and regression models and find the regression tasks were more challenging, with the best performing models achieving an R^2 coefficient of only 0.337. This shows that, while prediction of the type of emotional response can be achieved with a reasonable degree of accuracy, prediction of the *magnitude* of emotional response that will be evoked based on the content of the tweet presents significant difficulty.

We conduct feature evaluation to understand the content, user, and sentiment factors of a tweet that are predictive of certain sentiments and emotions in replies. Finally, we perform error analysis of the model results, analyzing cases of unexpected reply affects that our model fails to predict based on the original tweets with a focus on unexpectedly negative replies. Based on this analysis we identify several factors which drive reply affects to be unpredictable by the model including the presence of abusive or harassing tweet responses. This qualitative demonstration that such adverse behaviors can fall outside the normal or expected response patterns serves as the first step to leveraging the sentiment dynamics of social media conversations as a predictive flag for the detection of targeted abuse or harassment. Future work will focus on enrichment of the feature space by incorporating image signals, exploration of the dynamics of tweet-reply sentiment over time, and application of response affect prediction to the detection of abusive and harassing behavior on social media.