Table 3: Accuracy of classifiers on the emoji prediction task. d refers to the dimensionality of each LSTM layer. Parameters are in millions.

	Params	Top 1	Top 5
Random	–	1.6%	7.8%
fasttext	12.8	12.8%	36.2%
DeepMoji (d = 512)	15.5	16.7%	43.3%
DeepMoji (d = 1024)	22.4	17.0%	43.8%

sions for this fastText classifier, thereby making it almost identical to only using the embedding layer from the DeepMoji model. The difference in top 5 accuracy between the fastText classifier (36.2%) and the largest DeepMoji model (43.8%) underlines the difficulty of the emoji prediction task. As the two classifiers only differ in that the DeepMoji model has LSTM layers and an attention layer between the embedding and Softmax layer, this difference in accuracy demonstrates the importance of capturing the context of each word.

4.2 Benchmarking

We benchmark our method on 3 different NLP tasks using 8 datasets across 5 domains. To make for a fair comparison, we compare DeepMoji to other methods that also utilize external data sources in addition to the benchmark dataset. An averaged F1-measure across classes is used for evaluation in emotion analysis and sarcasm detection as these consist of unbalanced datasets while sentiment datasets are evaluated using accuracy.

An issue with many of the benchmark datasets is data scarcity, which is particularly problematic within emotion analysis. Many recent papers proposing new methods for emotion analysis such as (Staiano and Guerini, 2014) only evaluate performance on a single benchmark dataset, SemEval 2007 Task 14, that contains 1250 observations. Recently, criticism has been raised concerning the use of correlation with continuous ratings as a measure (Buechel and Hahn, 2016), making only the somewhat limited binary evaluation possible. We only evaluate the emotions {Fear, Joy, Sadness} as the remaining emotions occur in less than 5% of the observations.

To fully evaluate our method on emotion analysis against the current methods we thus make use of two other datasets: A dataset of emotions in tweets related to the Olympic Games created by Sintsova et al. that we convert to a single-label classification task and a dataset of self-reported emotional experiences created by a large group of psychologists (Wallbott and Scherer, 1986). See the supplementary material for details on the datasets and the preprocessing. As these two datasets do not have prior evaluations, we evaluate against a state-of-the-art approach, which is based on a valence-arousal-dominance framework (Buechel and Hahn, 2016). The scores extracted using this approach are mapped to the classes in the datasets using a logistic regression with parameter optimization using cross-validation. We release our preprocessing code and hope that these 2 two datasets will be used for future benchmarking within emotion analysis.

We evaluate sentiment analysis performance on three benchmark datasets. These small datasets are chosen to emphasize the importance of the transfer learning ability of the evaluated models. Two of the datasets are from SentiStrength (Thelwall et al., 2010), SS-Twitter and SS-Youtube, and follow the relabeling described in (Saif et al., 2013) to make the labels binary. The third dataset is from SemEval 2016 Task4A (Nakov et al., 2016). Due to tweets being deleted from Twitter, the SemEval dataset suffers from data decay, making it difficult to compare results across papers. At the time of writing, roughly 15% of the training dataset for SemEval 2016 Task 4A was impossible to obtain. We choose not to use review datasets for sentiment benchmarking as these datasets contain so many words pr. observation that even bag-of-words classifiers and unsupervised approaches can obtain a high accuracy (Joulin et al., 2016; Radford et al., 2017).

The current state of the art for sentiment analysis on social media (and winner of SemEval 2016 Task 4A) uses an ensemble of convolutional neural networks that are pretrained on a private dataset of tweets with emoticons, making it difficult to replicate (Deriu et al., 2016). Instead we pretrain a model with the hyperparameters of the largest model in their ensemble on the positive/negative emoticon dataset from Go et al. (2009). Using this pretraining as an initialization we finetune the model on the target tasks using early stopping on a validation set to determine the amount of training. We also implemented the Sentiment-Specific Word Embedding (SSWE) using the embeddings available on the authors' website (Tang et al., 2014), but found that it performed worse