Bidirectional Dilated LSTM with Attention for Fine-grained Emotion Classification in Tweets

Annika M Schoene[0000−0002−9248−617X], Alexander P Turner[0000−0002−2392−6549], and Nina Dethlefs[0000−0002−6917−5066]

The University of Hull, Cottingham Road, Hull HU6 7RX
amschoene@gmail.com

Abstract. We propose a novel approach for fine-grained emotion classification in tweets using a Bidirectional Dilated LSTM (BiDLSTM) with attention. Conventional LSTM architectures can face problems when classifying long sequences, which is problematic for tweets, where crucial information is often attached to the end of a sequence, e.g. an emoticon. We show that by adding a bidirectional layer, dilations and attention mechanism to a standard LSTM, our model overcomes these problems and is able to maintain complex data dependencies over time. We present experiments with two datasets, the 2018 WASSA Implicit Emotions Shared Task and a new dataset of 240,000 tweets. Our BiDLSTM with attention achieves a test accuracy of up to 81.97% outperforming competitive baselines by up to 10.52% on both datasets. Finally, we evaluate our data against a human benchmark on the same task.

Keywords: Natural Language Processing · Sentiment Analysis· Recurrent Neural Networks.

1 Introduction

There has been a surge of interest in the field of sentiment analysis in recent years, which is likely due to the growing number of social media users, who increasingly express their opinions, beliefs and attitudes in online posts towards a range of different topics, events and products [37]. Most sentiment analysis approaches to date focus on polarity detection [17, 3] but neglect the classification of more fine-grained emotion categories, such as Ekman's basic six emotions [15]. Fine-grained emotion detection has promising applicability in a number of domains, including detecting cyber-bullying [55] or identifying potential mental health issues in social media posts [44].

The majority of current approaches to sentiment analysis rely on deep learning algorithms [59], such as recurrent neural networks (RNN) [47, 25] and convolutional neural networks (CNN) [12, 11]. While tweets have previously been categorised as short sequences or sentence-level sentiment analysis [22], we argue that this should no longer be the case especially since Twitter increased its allowed character limit from 140 to 280 [49]. As such, tweets mostly face

 In: N. Chhaya, K. Jaidka, J. Healey, L. H. Ungar, A. Sinha (eds.): Proceedings of the 3rd Workshop of Affective Content Analysis, New York, USA, 07-FEB-2020, published at http://ceur-ws.org

also problems with classifying long sequences, similar to other natural language processing tasks [20].

In this paper we propose the use of Dilated RNNs (DRNN) for emotion classification from tweets. DRNNs introduce skip connections into a standard RNN to increase the range of temporal dependencies that can be modelled. Experiments on sequence classification for language modelling on the Penn Treebank, pixel-by-pixel MNIST classification and speaker identification from audio [10] have shown to outperform competitive baselines such as standard LSTM/GRU architectures as well as more specialised models. We expect that the same advantages can be observed for tweets. We extend the standard proposed DRNN with an embedding layer, bidirectional layer and attention mechanism and apply it to the classification of six basic emotion categories, *anger, fear, disgust, surprise, joy* and *sadness*. Figure 1 shows an example of a tweet.

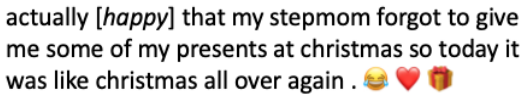

Fig. 1. Example of a tweet from the *'Joy'* category.

Therefore we hypothesise that by using dilated recurrent neural networks we can take advantage of the increased sequence length of tweets and avoid information loss over time. Another reason for the good performance of dilated recurrent skip connections is that they have a better balance of memory over a larger period of time compared to standard RNNs. We believe that using a similar structure, albeit not for a very long sequence but treating tweets as longer sequence will enable us to achieve better classification accuracies compared to treating tweets as a short sequence problem.

We experiment with two datasets, the 2018 WASSA Implicit Emotions Shared Task dataset which contains 153,383 tweets and can be considered an established benchmark. In addition, we collected a new larger dataset of 240,000 tweets using the same six emotion categories. We find that on both datasets, DLSTMs with attention perform better than standard LSTM or CNN architectures, as well as any of the submissions to the WASSA shared task, achieving up to 71.45% of accuracy. We find that the BiDLSTMs with attention are particularly beneficial for the longest sequences in our datasets and that the additions of a word embeddings, bidirectional layer and attention mechanism further increase performance.

2 Related Work

Recently, deep learning methods for sentiment and emotion classification have become the predominant technique. For example, [23] developed a soft attention-based LSTM with CNN for sarcasm detection. Work conducted by [36] use a

deep CNN with a multi-kernel classifier to extract features of short sequences for multi-modal sentiment analysis and show that this increases accuracy. [42] use a BiLSTM for a range of different text classification tasks, including sentiment analysis. In their experiments they show that using a single-layer BiLSTM with pretrained word embeddings and trained with cross-entropy loss achieves competitive results compared to more complex learning models. Most recently the Implicit Emotions Shared Task (IEST) [21] has used Tweets, where the winning model, named *'Amobee'*, was able to outperform the baseline score significantly by achieving an accuracy of 71.45% [41]. Amobee is a bidirectional GRU with an additional attention mechanism inspired by [5] and additional hidden layers. It has been reasoned that the model's success has been due to its specific type of transfer learning. The baseline model for this shared task was established using a maxentropy classifier with L2 regularization, where the F1 score reached an accuracy of 59.1% on the test data. Recurrent neural networks have become the predominant neural network across as range of sentiment analysis and emotion detection tasks [13]. Similarly, almost half of the submissions to the annual SemEval shared task [39, 27, 29] used some form of neural networks. At the same time, the majority of approaches to detect sentiment continue to focus on polarity detection [9], including approaches to identifying sentiment on social media such as Twitter [39, 30] or longer texts such as reviews or blogs [32]. This is limiting for real-world applications, where for mental state detection, customer reviews, advertising, and many more, fine-grained emotions can add substantial added value.

Approaches that have attempted more fine-grained classification are mostly based on Ekman's six basic emotions [15], *anger, fear, disgust, surpise, joy* and *sadness*, or Plutchik's eight basic emotions [35], who extended [15] basic emotions with *Trust* and *Anticipation*. For example, [2] apply Gated Recurrent Neural Networks (GRNNs) to classify tweets collected based on hashtags carrying emotions into [35] emotion categories. Research conducted by [28] used hashtags that contain emotion words based on Plutschnik's eight basic emotions to show that user-labelled hashtags used as annotations are consistent with those annotated by trained judges. Furthermore a new lexicon based on the same twitter corpus is introduced. [26] introduces a Topic Sentiment Model (TSM), which can capture both topics and sentiment. The model is based on Probabilistic Latent Semantic Indexing (pLSI) and utilises an online sentiment retrieval service to induce prior knowledge to the model. Research by [46] use distant supervision and a lexicon to label tweets for Plutschik's eight basic emotions [35] and then classify them. Work conducted by [40] also investigated eight basic emotions in online discourse. [8] used the whole taxonomy of Plutschik's emotions to analyse chat messages.

Work on sentiment classification from social media has additionally explored the occurrence of emoticons and their influence on sentiment classification [16]. [31] conducted research distinguishing happiness and sadness in emoticons. Similarly, [22] have shown that the usage of both hashtags and emoticons can be beneficial and contribute to more accurate classification of tweets.

3 Learning Model

Motivation There are a number of challenges that have to be taken into account when using recurrent neural networks to learn longer sequences, which include but are not limited to: (1) maintaining mid- and short term memory is problematic when memorising long-term dependencies [19] and (2) vanishing and exploding gradient descent [33]. Therefore it could be argued that there is a need for a more specialised learning model which can overcomes these challenges. [52] introduce a dilated LSTM as part of a reinforcement learning task, where the learning model has one dilated recurrent layer with fixed dilations. Work by [10] introduced a Dilated RNN by using dilated skip connections. The dilated LSTM alleviates the problem of learning long sequences, however not every word in a sequence has the same meaning or importance. Therefore we extend this network by (1) an embedding layer, (2) a bidirectional layer and (3) attention mechanism. The full architecture of the Bidirectional Dilated LSTM (BiDLSTM) with attention is shown in Figure 2.

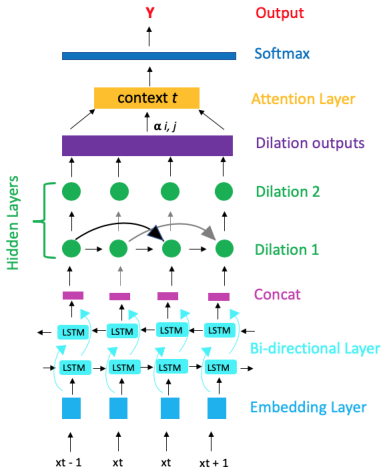

Fig. 2. bidirectional DLSTM with attention

LSTM architecture Our primary model is the Long-short-term memory (LSTM) given its suitability for language and time-series data [20]. We feed into the LSTM an input sequence $\mathbf{x} = (x_1, \ldots, x_N)$ of words in a tweet alongside a label $y \in Y$ denoting an emotion from any of the six basic emotion categories. The LSTM learns to map inputs x to an output y via a hidden representation $\mathbf{h}_t$ which can be found recursively from an activation function:

$$f(\mathbf{h}_{t-1}, x_t), \tag{1}$$

where t denotes a time-step. During training, we minimise a loss function, in our case categorical cross-entropy, as:

$$L(x, y) = -\frac{1}{N} \sum_{n \in N} x_n \log y_n. \tag{2}$$

Standard LSTMs manage their weight updates through a number of gates that determine the amount of information that should be retained and forgotten at each time step. In particular, we distinguish an 'input gate' i that decides how much new information to add at each time-step, a 'forget gate' f that decides what information not to retain and an 'output gate' o determining the output. More formally, and following the definition by [18], this leads us to update our hidden state $\mathbf{h}$ as follows (where σ refers to the logistic sigmoid function, c is the 'cell state', W is the weight matrix and b is the bias term):

$$i_t = \sigma(W_{xi} x_t + W_{hi} h_{t-1} + W_{ci} c_{t-1} + b_i) \tag{3}$$

$$f_t = \sigma(W_{xf} x_t + W_{hf} h_{t-1} + W_{cf} c_{t-1} + b_f) \tag{4}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc} x_t + W_{hc} h_{t-1} + b_c) \tag{5}$$

$$o_t = \sigma(W_{xo} x_t + W_{ho} h_{t-1} + W_{co} c_t + b_o) \tag{6}$$

$$h_t = o_t \tanh(c_t) \tag{7}$$

A standard LSTM definition solves some of the problems of vanilla RNNs have, such as the vanishing gradient descent problem [20], but it still has some shortcomings when learning long-term dependencies. One of them is due to the cell state of an LSTM; the cell state is changed by adding some function of the inputs. When we backpropagate and take the derivative of c_t with respect to $c_t - 1$, the added term would disappear and less information would travel through the layers of a learning model. This shortcoming can be addressed through the use of dilations and skip-connections in the dilated LSTM.

Embedding and bidirectional Layer Each tweet t contains w_i words where $w_i t, t \in [0, T]$ represents the ith word in each tweet. We utilise GloVe word embeddings trained on 2 billion tweets as developed by [34], in our 200-dimensional embedding layer. Then we use a bidirectional LSTM to obtain information from both directions of each word in order to capture the contextual information. The bidirectional LSTM incorporates the forward LSTM $\overrightarrow{h}_{t^{(i)}}$ which reads each tweet from $w_i 1$ to $w_i T$ and a backward LSTM $\overleftarrow{h}_{t^{(i)}}$ which reads words in each tweet from $w_i T$ to w_i, where $x_i t$ represents word vectors in an embedding matrix:

$$\mathbf{x_i t} = W_e w_i t, t \in [1, T] \tag{8}$$

$$\overrightarrow{h}_{t^{(i)}} = \overrightarrow{LSTM}(\mathbf{x_i t}), t \in [1, T] \tag{9}$$

$$\overleftarrow{h}_{t^{(i)}} = \overleftarrow{LSTM}(\mathbf{x_i t}), t \in [1, T] \tag{10}$$

We then concatenate all outputs of the forward hidden state $\overrightarrow{h}_t$ and backward hidden state $\overleftarrow{h}_t$, where the output o allows us to utilise all information available in each tweet. The output o is then fed into the Dilated LSTM.

Dilated LSTM Layer For our implementation of a Dilated LSTM, we follow the implementation of recurrent skip connections with exponentially increasing dilations in a multi-layered learning model - as proposed by [10] - as it allows LSTMs to better learn input sequences and their dependencies. This means that temporal and complex data dependencies are learned on different layers. The most important part of this architecture is the dilated recurrent skip connection in the LSTM cell, where $c_t^{(l)}$ is the cell in layer l at time t:

$$c_t^{(l)} = LSTM(\mathbf{o}_t^{(l)}, c_{t-s^l}^{(l)}) \cdot \tag{11}$$

$s^{(l)}$ is the skip length of layer l; $\mathbf{o}_t^{(l)}$ as the input to layer l at time t in a LSTM. The exponentially increasing dilations across layers have been inspired by [51]; $s^{(l)}$ denotes the dilation of the l-th layer, where M and L denotes dilations at different layers:

$$s^{(l)} = M^{(l-1)}, l = 1, \dots L. \tag{12}$$

As outlined by [10] there are two main benefits to stacking exponentially dilated recurrent layers: (1) it enables different layers to focus on different temporal resolutions and (2) it reduces the length of paths between nodes at different time-steps, which enables the network to learn more complex long-term dependencies. Therefore exponentially increasing dilations shortens any given sequence length at different layers.

Attention Layer The attention mechanism was first introduced by [4], but has since been used in a number of different tasks including machine translation [24], sentence pairs detection [58], neural image captioning [56] and action recognition [45].

Our implementation of the attention mechanism is inspired by [57], using attention to find words that are most important to the meaning of a tweet. We use the output of the dilated LSTM as direct input into the attention layer, where O denotes the output of final layer L of the Dilated LSTM at time t_{+1}. The *attention* for each word w in a tweet t is computed as follows, where h_{iw} is the hidden representation of the dilated LSTM output, α_{iw} represents normalised alpha weights measuring the importance of each word and t_i is the corresponding tweet vector:

$$u_{iw} = \tanh(\mathbf{O} + b_w) \tag{13}$$

$$\alpha_{iw} = \frac{\exp\left(h_{iw}^T\right)}{\sum_t \exp\left(h_{iw}^T\right)} \tag{14}$$

$$\mathbf{t_i} = \sum_t \alpha_{iw} o \cdot \tag{15}$$

4 Experiments

We present the datasets used, our baselines and discuss objective and subjective results.

4.1 Data

We will work with the following datasets:

- The WASSA Implicit Emotions Shared Task (IEST) [21] data consists of 155,383 tweets and is based on [15] six basic emotions.
- The Ekman's Emotion Keyword (EEK) data, a collection of 240,000 tweets that we collected between September 2017 and December 2018. [1]

Table 1 shows a comparison of the two datasets in terms of their size and basic distribution of emotion categories represented in them.

Emotion	IEST	EEK
Anger	25,384	40,000
Fear	25,387	40,000
Disgust	25,396	40,000
Surprise	25,402	40,000
Joy	25,377	40,000
Sadness	25,396	40,000

Table 1. Comparison of IEST and EEK dataset emotion category distribution

Emotion	Keywords
Anger	anger,angry, furious
Fear	fear, scared, fearful
Disgust	disgust, disgusting
Surprise	surprise, surprising
Joy	joy, happy
Sadness	sad

Table 2. Synonyms for Twitter API queries

Both datasets were collected using the Twitter API [50] and a list of keyword and synonyms were specified for automatic data collection from Twitter. See Table 2 for the keywords that we used, following [21] and using Ekman's six basic emotions. After the initial data collection we filtered tweets by those marked in the language tab as "English" and removed any duplicates. Then we used the text processing library developed by [6], to anonymise usernames and mask URLs. Afterwards we used a dictionary containing all emotion keywords listed in Table 2 and replaced existing keywords in all tweets with the term *[keyword]*. Finally each tweet was assigned a label based on the emotion category its keyword belonged to (see Figure 1). For our experiments we use 80% of the data for training, 10% for validation and the remaining 10% for testing.

[1] The dataset will be released to the research community upon request and in accordance with the Twitter API guidelines [50]

4.2 Baselines

Similarly to [21] we use a a maximum entropy classifier with L2 regularisation for establishing the baselines of our datasets. All baselines will be evaluated in two conditions:

Capped length , where we cap the length of any sequence to 40 in accordance with the WASSA IEST challenge winners.

Full length ,where we use the average full uncapped length of a sequence (maximum 103). Our intuition is that this condition will particularly reveal the advantages of the skip connections.

For the DLSTM, BiDLSTM and BiDLSTM with attention, we established the number of dilations empirically. There are two dilated layers with the dilations increasing exponentially starting at *1 [1,2]*. This means that each sub-LSTM for the pruned sequence has the following sequence length *[Dilation 1 = 40, Dilation 2 = 20]* with a total of 20 hidden units per layer. Whilst each sub-LSTM for the longer sequence has the following sequence length:*[Dilation 1 = 102, Dilation 2 = 51]*.
We evaluate our BiDLSTM with attention against the following baselines:

- **DLSTM** – a dilated LSTM with hierarchically stacked dilations and hyper-parameters: learning rate: 0.001, batch size: 128, optimizer: Adam, dropout: 0.5
- **BiDLSTM** – a two-layer bidirectional dilated LSTM with a three-layer LSTM, hierarchically stacked dilations and the same hyperparameters as the DLSTM.
- **BiLSTM** – a BiLSTM with 2 layers and the following hyper-parameters: learning rate: 0.001, batch size: 128, optimizer: Adam, dropout: 0.5. This model is similar to recent work by [42] who used a single layer biLSTM to classify the ImdB movie review dataset into positive and negative reviews.
- **BiLSTM with attention** – a BiLSTM with attention and the following hyper-parameters: learning rate: 0.001, batch size: 128, optimizer: Adam, dropout: 0.5. This model is similar to recent work by [7, 43].
- **CNN** – a CNN 2-D convolution with two fully connected layers, a filter size of *1,2* and *102* filters, and a ReLU function. This learning model is similar to recent work by [14].
- **CNN-LSTM** – we follow the implementation of the learning model by [53], using a CNN that is feeding into an LSTM. This model was used to predict the valence/arousal of ratings in textual data.

Also, we compare our model against the winner of the 2019 WASSA IEST dataset, called Amobee[41]. All of the experiments conducted using Tensorflow [1].

5 Results

We benchmark the BiDLSTM with attention to a number of different neural networks, using both vanilla neural networks and more specialised neural networks that have been used in sentiment analysis tasks. We compare results by two different sequence length and use four different metrics for evaluation; test set accuracy, precision, recall and F1-score.

Capped Sequences Tables 3 and 4 show the results for capped sequences lengths for both the IEST and EEK dataset respectively.

Learning Model	Test Acc.	Precision	Recall	F1-score
Max Entropy	58.4	0.59	0.57	0.58
CNN	43.17	0.44	0.42	0.43
CNN LSTM	55.42	0.56	0.54	0.55
BI LSTM	49.47	0.50	0.48	0.49
BI LSTM attention	58.60	0.60	0.56	0.58
DLSTM	56.44	0.57	0.55	0.56
BiDLSTM	67.96	0.68	0.67	0.67
Amobee	-	-	-	**71.45**
BiDLSTM attention	**72.83**	**0.74**	**0.71**	**0.72**

Table 3. Results for capped sequences (IEST Dataset)

Learning Model	Test Acc.	Precision	Recall	F1-score
Max Entropy	62.50	0.63	0.62	0.62
CNN	55.33	0.56	0.54	0.55
CNN LSTM	59.79	0.60	0.59	0.59
BI LSTM	60.19	0.61	0.59	0.60
BI LSTM attention	63.62	0.64	0.62	0.63
DLSTM	66.80	0.67	0.65	0.66
BiDLSTM	69.71	0.70	0.69	0.69
BiDLSTM attention	**73.74**	**0.75**	**0.72**	**0.73**

Table 4. Results for capped sequences (EEK dataset)

It can be seen that vanilla CNN and BiLSTM fall just short of the baselines established for this task. The CNN-LSTM and DLSTM architecture, both outperform their vanilla predecessors. The BiLSTM with attention and BiDLSTM surpass the baselines but falls short of the model proposed in the IEST task for both datasets. It can be seen that BiDLSTM with attention outperforms all previous models on the capped sequence length by over 14.43% for capped sequences and the IEST baseline by 11.24%. The results for capped sequence length using the IEST dataset (Table 3) show that our proposed model surpasses the *'Amobee'* model's result, however this is only marginally. We hypothesis that the reason

the DLSTM, BiDLSTM and BiDLSTM and with attention either fall short of the baselines or only marginally surpass them is due the model not being able to take full advantage of the full sequence length.

Long sequences Table 5 shows the results for the IEST dataset using full length sequences and Table 6 also shows the results for the full length for the EEK dataset. Similarly to the results for the capped sequence length, the CNN and Bi-LSTM fall short of the established baselines. Only the CNN-LSTM improves the performance of the results, whereas for the long sequences the DLSTM, BiLSTM with attention and BiDSLTM surpasses the baselines of both datasets. The BiDLSTM with attention outperforms all models on the full length sequences by over 20.36% on the EEK dataset and the IEST baseline by 18.47%. These results show that incorporating contextual information through the bidirectional layer and using attention to focus on the most important words in a tweet enhances the dilated LSTMs ability to cope with longer sequences. This confirms that using more specialised learning models such as the DLSTM, BiDLSTM and BiDLSTM with attention allow us to better capture information in longer sequences.

Learning Model	Test Acc.	Precision	Recall	F1-score
Max Entropy	58.4	0.59	0.57	0.58
CNN	43.95	0.44	0.43	0.43
CNN LSTM	56.15	0.57	0.55	0.56
BI LSTM	51.73	0.52	0.51	0.51
BI LSTM attention	58.79	0.59	0.58	0.58
DLSTM	60.27	0.61	0.59	0.60
BiDLSTM	69.01	0.71	0.67	0.69
BiDLSTM attention	**78.76**	**0.79**	**0.78**	**0.78**

Table 5. Results for full length (IEST dataset)

Learning Model	Test Acc.	Precision	Recall	F1-score
Max Entropy	62.50	0.63	0.62	0.62
CNN	55.12	0.56	0.54	0.55
CNN LSTM	60.11	0.61	0.59	0.60
BI LSTM	60.88	0.61	0.60	0.60
BI LSTM attention	62.70	0.63	0.62	0.62
DLSTM	67.18	0.68	0.66	0.67
BiDLSTM	69.53	0.71	0.68	0.69
BiDLSTM attention	**80.97**	**0.82**	**0.79**	**0.80**

Table 6. Results for full length (EEK dataset)

5.1 Evaluation of Prediction Labels

In order to evaluate the performance of each model, we have set aside 5,000 tweets per dataset that have not been used during training or testing previously. We then use the pretrained models to establish, which labels are hardest to predict for each network. We compare the best performing learning model with human performance. For this we used Amazon Mechanical Turk [48], where each tweet was annotated by three different annotators for the six emotion categories, yielding 15,000 annotations per dataset. All emotion words were replaced with the term *'[Keyword]'*, a sample tweet can be seen in Figure 3.

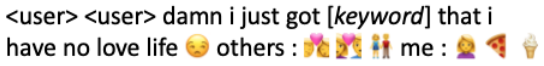

Fig. 3. Example of a tweet shown to annotators.

We use confusion matrices to visualise the quality of label output for our learning model on both datasets. Figures 4 and 5 both show the confusion matrices for the BiDLSTM with attention. Figures 4 and 5 shows that for the both datasets *Joy* was most accurately predicted emotion, whilst *Anger* (61.96 %) was often misclassified. Furthermore it is shows that *Anger* is more often confused with *Disgust* in both datasets.

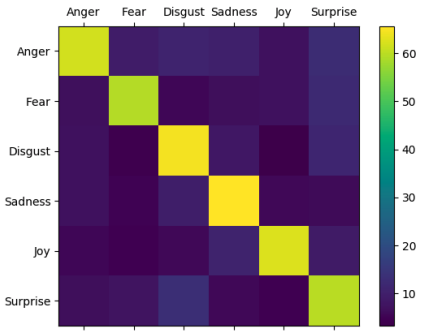

Fig. 4. BiDLSTM attention (IEST)

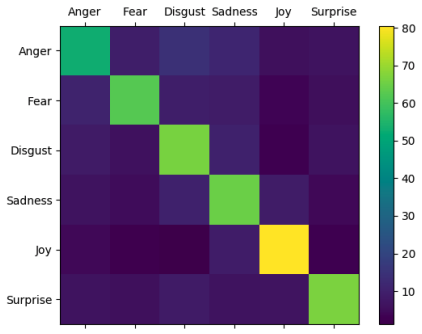

Fig. 5. BiDLSTM attention (EEK)

Furthermore we have also looked at each emotion in both datasets in order to gain a better insight into how well each emotion is classified by the proposed learning model. We use Precision, Recall and F-1 score as our evaluation metrics for both of the test datasets. Table 7 shows the emotion labels in the IEST dataset using the full sequence length, where the best performing emotion is *Joy* and the emotion *Anger* is most often misclassfied. Table 8 also shows the

label classification for the EEK dataset using the full sequence length, confirming that the same emotions, *Joy* and *Anger*, are also the most and least likely to be accurately classified.

Type	Precision	Recall	F1-score
Anger	**0.69**	**0.76**	**0.72**
Fear	0.69	0.83	0.75
Disgust	0.83	0.75	0.79
Sadness	0.76	0.78	0.77
Joy	**0.90**	**0.75**	**0.82**
Surprise	0.84	0.78	0.81
Average	0.79	0.78	0.78

Table 7. Evaluation metrics per emotion label - BiDLSMT with attention in % (IEST dataset)

Type	Precision	Recall	F1-score
Anger	**0.71**	**0.79**	**0.75**
Fear	0.74	0.86	0.80
Disgust	0.84	0.78	0.81
Sadness	0.78	0.81	0.79
Joy	**0.93**	**0.77**	**0.85**
Surprise	0.84	0.79	0.81
Average	0.81	0.80	0.80

Table 8. Evaluation metrics per emotion label - BiDLSMT with attention in % (EEK dataset)

Afterwards we looked at the results for the human annotation, for the same test datasets. Figures 6 and 7 show the confusion matrices for the human annotators. Each confusion matrix shows the number of correctly and false predicted labels in percentages. We have found that for both datasets evaluated by humans that the most commonly correctly annotated emotion was *Joy* with 37.70% in the IEST and 41.80% in the EEK dataset. The emotion *Disgust* was least likely to be accurately annotated in both datasets. Furthermore *Disgust* was most often mistaken for the emotion *Sadness* in both datasets and overall there were far fewer accurately predicted labels by the human annotators compared to the proposed learning model.

Fig. 6. Humans annotators (IEST)

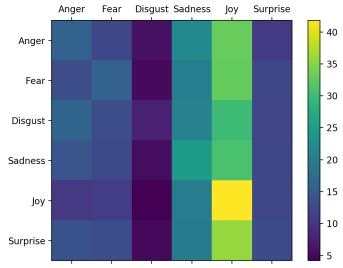

Fig. 7. Humans annotators (EEK)

In Figure 8 we show an example of a tweet with its true label and the labels predicted by human annotators. It can be seen that for all three people anno-

tating this tweet there was no agreement on the emotion label and no annotator picked the correct label. This illustrate how hard this task may be for humans as the keyword could have been replaced with a number of different emotion keywords and made sense.

"that one girl from my art class said she feels [keyword] when she sees children and pregnant women"

True Label: *Disgust*
Predicted Label: *Surprise, Fear and Sadness*

Fig. 8. A tweet illustrating the difficulty of the task for a human annotator to choose one emotion keyword.

Probabilities of labels Furthermore we have looked at 100 random test samples to see the probability distribution of the output labels (see Figures 9 and 10). It could be argued that there might be some larger pattern that is detected by learning models when humans write about emotion that may not be detected by humans on a qualitative basis.

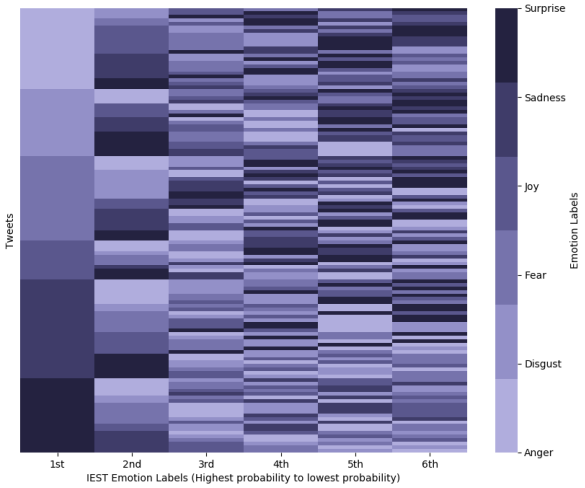

Fig. 9. Visualisation of IEST Emotion labels based on the probability of accurate prediction - BiDLSTM with attention

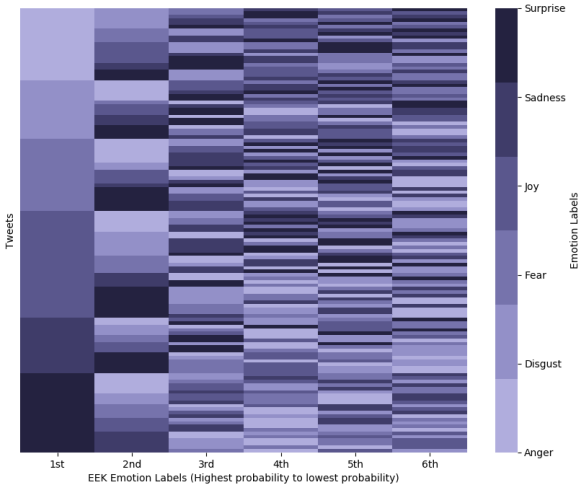

Fig. 10. Visualisation of EEK Emotion labels based on the probability of accurate prediction - BiDLSTM with attention

This might be due to the difficulty in the task where many emotions are closely related or overlapping such as *Disgust* and *Anger*, where humans were not able to interpret them correctly [54]. Other studies have previously found that humans struggle to identify emotions in textual data due to the lack of extra information provided (e.g.: tone of voice or facial expression) and therefore often projecting their own emotional state and information [38]. However, this is not possible for any learning model and therefore might be the reason why they are better at detecting underlying patterns in this type of data.

6 Conclusion

In this paper we have found that our learning model, the bidirectional dilated LSTM with attention, performs above the baseline of 58.4% by over 14.43% on the WASSA shared task dataset. Furthermore, our model performs also best on our own dataset achieving an accuracy of 80.97%. We have also found that when using longer sequences we achieve better results with models that are more specialised compared to vanilla neural networks. Additionally, we have shown that when pruning our model to use a shorter input sequence it still outperforms state-of-the art results. Also, it could be argued that treating tweets as longer sequences we can utilise more information in a tweet. Furthermore we have evaluated which labels are most likely predicted correctly by both humans and the BiDLSTM with attention. We have demonstrated that the task of accurately

identifying the six emotion categories in tweets is considerably harder for humans compared to the learning model. This could largely be due to the amount of emotions projected by humans on an individual tweet which doesn't enable them to identified overall patterns on a qualitative basis. Also, we have outlined the collection of a new resource, a dataset of 240,000 tweets that have been labelled for six emotion categories.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: a system for large-scale machine learning. In: OSDI. vol. 16, pp. 265–283 (2016)
2. Abdul-Mageed, M., Ungar, L.: Emonet: Fine-grained emotion detection with gated recurrent neural networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 718–728 (2017)
3. Amplayo, R.K., Kim, J., Sung, S., Hwang, S.w.: Cold-start aware user and product attention for sentiment classification. arXiv preprint arXiv:1806.05507 (2018)
4. Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate. In: Proc. of the International Conference on Learning Representations (ICLR). San Diego, CA, USA (2015)
5. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
6. Baziotis, C., Pelekis, N., Doulkeridis, C.: Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 747–754. Association for Computational Linguistics, Vancouver, Canada (August 2017)
7. Baziotis, C., Pelekis, N., Doulkeridis, C.: Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 747–754 (2017)
8. Brooks, M., Kuksenok, K., Torkildson, M.K., Perry, D., Robinson, J.J., Scott, T.J., Anicello, O., Zukowski, A., Harris, P., Aragon, C.R.: Statistical affect detection in collaborative chat. In: Proceedings of the 2013 conference on Computer supported cooperative work. pp. 317–328. ACM (2013)
9. Cambria, E.: Affective computing and sentiment analysis. IEEE Intelligent Systems **31**(2), 102–107 (2016)
10. Chang, S., Zhang, Y., Han, W., Yu, M., Guo, X., Tan, W., Cui, X., Witbrock, M., Hasegawa-Johnson, M.A., Huang, T.S.: Dilated recurrent neural networks. In: Advances in Neural Information Processing Systems. pp. 77–87 (2017)
11. Chen, T., Xu, R., He, Y., Wang, X.: Improving sentiment analysis via sentence type classification using bilstm-crf and cnn. Expert Systems with Applications **72**, 221–230 (2017)
12. Dahou, A., Elaziz, M.A., Zhou, J., Xiong, S.: Arabic sentiment classification using convolutional neural network and differential evolution algorithm. Computational Intelligence and Neuroscience **2019** (2019)
13. Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Hoi, G.W.S., Zubiaga, A.: Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. arXiv preprint arXiv:1704.05972 (2017)

14. Dos Santos, C., Gatti, M.: Deep convolutional neural networks for sentiment analysis of short texts. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 69–78 (2014)
15. Ekman, P., Levenson, R.W., Friesen, W.V.: Autonomic nervous system activity distinguishes among emotions. Science **221**(4616), 1208–1210 (1983)
16. Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., Lehmann, S.: Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. arXiv preprint arXiv:1708.00524 (2017)
17. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford **1**(12) (2009)
18. Graves, A.: Generating Sequences With Recurrent Neural Networks. CoRR **abs/1308.0850** (2013), http://arxiv.org/abs/1308.0850
19. Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J., et al.: Gradient flow in recurrent nets: the difficulty of learning long-term dependencies (2001)
20. Hochreiter, S., Schmidhuber, J.: Lstm can solve hard long time lag problems. In: Advances in neural information processing systems. pp. 473–479 (1997)
21. Klinger, R., De Clercq, O., Mohammad, S.M., Balahur, A.: Iest: Wassa-2018 implicit emotions shared task. arXiv preprint arXiv:1809.01083 (2018)
22. Kouloumpis, E., Wilson, T., Moore, J.D.: Twitter sentiment analysis: The good the bad and the omg! Icwsm **11**, 164 (2011)
23. Kumar, A., Sangwan, S.R., Arora, A., Nayyar, A., Abdel-Basset, M., et al.: Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network. IEEE Access (2019)
24. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
25. Ma, Y., Peng, H., Cambria, E.: Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
26. Mei, Q., Ling, X., Wondra, M., Su, H., Zhai, C.: Topic sentiment mixture: modeling facets and opinions in weblogs. In: Proceedings of the 16th international conference on World Wide Web. pp. 171–180. ACM (2007)
27. Mohammad, S., Bravo-Marquez, F., Salameh, M., Kiritchenko, S.: Semeval-2018 task 1: Affect in tweets. In: Proceedings of The 12th International Workshop on Semantic Evaluation. pp. 1–17 (2018)
28. Mohammad, S.M.: # emotional tweets. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation. pp. 246–255. Association for Computational Linguistics (2012)
29. Mohammad, S.M., Bravo-Marquez, F.: Wassa-2017 shared task on emotion intensity. arXiv preprint arXiv:1708.03700 (2017)
30. Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., Stoyanov, V.: Semeval-2016 task 4: Sentiment analysis in twitter. In: Proceedings of the 10th international workshop on semantic evaluation (semeval-2016). pp. 1–18 (2016)
31. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREc. vol. 10, pp. 1320–1326 (2010)
32. Pang, B., Lee, L., et al.: Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval **2**(1–2), 1–135 (2008)
33. Pascanu, R., Mikolov, T., Bengio, Y.: Understanding the exploding gradient problem. CoRR, abs/1211.5063 (2012)

34. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
35. Plutchik, R.: The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American scientist **89**(4), 344–350 (2001)
36. Poria, S., Cambria, E., Gelbukh, A.: Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 2539–2544 (2015)
37. Ravi, K., Ravi, V.: A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowledge-Based Systems **89**, 14–46 (2015)
38. Riordan, M.A., Trichtinger, L.A.: Overconfidence at the keyboard: Confidence and accuracy in interpreting affect in e-mail exchanges. Human Communication Research **43**(1), 1–24 (2017)
39. Rosenthal, S., Farra, N., Nakov, P.: Semeval-2017 task 4: Sentiment analysis in twitter. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 502–518 (2017)
40. Rothkrantz, L.: Online emotional facial expression dictionary. In: Proceedings of the 15th International Conference on Computer Systems and Technologies. pp. 116–123. ACM (2014)
41. Rozental, A., Fleischer, D.: Amobee at semeval-2018 task 1: Gru neural network with a cnn attention mechanism for sentiment classification. arXiv preprint arXiv:1804.04380 (2018)
42. Sachan, D.S., Zaheer, M., Salakhutdinov, R.: Revisiting lstm networks for semi-supervised text classification via mixed objective function (2018)
43. Schoene, A.M., Dethlefs, N.: Unsupervised suicide note classification (2018)
44. Schoene, A.M., Dethlefs, N.: Automatic identification of suicide notes from linguistic and sentiment features. In: Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. pp. 128–133 (2016)
45. Sharma, S., Kiros, R., Salakhutdinov, R.: Action recognition using visual attention. arXiv preprint arXiv:1511.04119 (2015)
46. Suttles, J., Ide, N.: Distant supervision for emotion classification with discrete binary values. In: International Conference on Intelligent Text Processing and Computational Linguistics. pp. 121–136. Springer (2013)
47. Tay, Y., Tuan, L.A., Hui, S.C.: Learning to attend via word-aspect associative fusion for aspect-based sentiment analysis. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
48. Turk, A.M.: Amazon mechanical turk. Retrieved August **17**, 2012 (2012)
49. Twitter: Counting characters. https://developer.twitter.com/en/docs/basics/counting-characters.html (Dec 2018), accessed on 2018-11-11
50. Twitter: Developer policy. https://developer.twitter.com/en.html (Dec 2018), accessed on 2018-11-11
51. Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A.W., Kavukcuoglu, K.: Wavenet: A generative model for raw audio. In: SSW. p. 125 (2016)
52. Vezhnevets, A.S., Osindero, S., Schaul, T., Heess, N., Jaderberg, M., Silver, D., Kavukcuoglu, K.: Feudal networks for hierarchical reinforcement learning. arXiv preprint arXiv:1703.01161 (2017)

53. Wang, J., Yu, L.C., Lai, K.R., Zhang, X.: Dimensional sentiment analysis using a regional cnn-lstm model. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 225–230 (2016)
54. Widen, S.C., Russell, J.A., Brooks, A.: Anger and disgust: Discrete or overlapping categories. In: 2004 APS Annual Convention, Boston College, Chicago, IL (2004)
55. Xu, J.M., Zhu, X., Bellmore, A.: Fast learning for sentiment analysis on bullying. In: Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining. p. 10. ACM (2012)
56. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057 (2015)
57. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1480–1489 (2016)
58. Yin, W., Schütze, H., Xiang, B., Zhou, B.: Abcnn: Attention-based convolutional neural network for modeling sentence pairs. Transactions of the Association for Computational Linguistics **4**, 259–272 (2016)
59. Young, T., Hazarika, D., Poria, S., Cambria, E.: Recent trends in deep learning based natural language processing. ieee Computational intelligenCe magazine **13**(3), 55–75 (2018)