Dilated LSTM with attention for Classification of Suicide Notes

Annika M Schoene George Lacey Alexander P Turner Nina Dethlefs
The University of Hull
Cottingham Road
Hull
HU6 7RX
amschoene@gmail.com

Abstract

In this paper we present a dilated LSTM with attention mechanism for document-level classification of suicide notes, last statements and depressed notes. We achieve an accuracy of 87.34% compared to competitive baselines of 80.35% (Logistic Model Tree) and 82.27% (Bi-directional LSTM with Attention). Furthermore, we provide an analysis of both the grammatical and thematic content of suicide notes, last statements and depressed notes. We find that the use of personal pronouns, cognitive processes and references to loved ones are most important. Finally, we show through visualisations of attention weights that the Dilated LSTM with attention is able to identify the same distinguishing features across documents as the linguistic analysis.

1 Introduction

Over recent years the use of social media platforms, such as blogging websites has become part of everyday life and there is increasing evidence emerging that social media can influence both suicide-related behaviour (Luxton et al., 2012) and other mental health conditions (Lin et al., 2016). Whilst there are efforts to tackle suicide and other mental health conditions online by social media platforms such as Facebook (Facebook, 2019), there are still concerns that there is not enough support and protection, especially for younger users (BBC, 2019). This has led to a notable increase in research of suicidal and depressed language usage (Coppersmith et al., 2015; Pestian et al., 2012) and subsequently triggered the development of new healthcare applications and methodologies that aid detection of concerning posts on social media platforms (Calvo et al., 2017). More recently, there has also been an increased use of deep learning techniques for such tasks (Schoene and Dethlefs, 2018), however there is little evidence which features are most relevant for the accurate classification. Therefore we firstly analyse the most important linguistic features in suicide notes, depressed notes and last statements. Last Statements have been of interest to researchers in both the legal and mental health community, because an inmates last statement is written, similarly to a suicide note, closely before their death (Texas Department of Criminal Justices, 2019). However, the main difference remains that unlike in cases of suicide, inmates on death row have no choice left in regards to when, how and where they will die. Furthermore there has been extensive analysis conducted on the mental health of death row inmates where depression was one of the most common mental illnesses. Work in suicide note identification has also compared the different states of mind of depressed and suicidal people, because depression is often related to suicide (Mind, 2013). Secondly, we introduce a recurrent neural network architecture that enables us to (1) model long sequences at document level and (2) visualise the most important words to accurate classification. Finally, we evaluate the results of the linguistic analysis against the results of the neural network visualisations and demonstrate how these features align. We believe that by exploring and comparing suicide notes with last statements and depressed notes, both qualitatively and quantitatively it could help us to find further differentiating factors and aid in identifying suicidal ideation.

2 Related Work

The analysis and classification of suicide notes, depression notes and last statements has traditionally been conducted separately. Work on suicide notes has often focused on identifying suicidal ideation online (O'dea et al., 2017) or distinguish-

Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis (LOUHI 2019), pages 136–145

https://doi.org/10.18653/v1/D19-62

ing genuine from forged suicide notes (Coulthard et al., 2016), whilst the main purpose of analysing last statements has been to identify psychological factors or key themes (Schuck and Ward, 2008).

Suicide Notes Recent years have seen an increase in the analysis of suicidal ideation on social media platforms, such as Twitter. Shahreen et al. (2018) searched the Twitter API for specific keywords and analysed the data using both traditional machine learning techniques as well as neural networks, achieving an accuracy of 97.6% using neural networks. Research conducted by Burnap et al. (2017) have developed a classifier to distinguish suicide-related themes such as the reports of suicides and casual references to suicide. Work by Just et al. (2017) used a dataset annotated for suicide risks by experts and a linguistic analysis tool (LIWC) to determine linguistic profiles of suicide-related twitter posts. Other work by Pestian et al. (2010) has looked into the analysis and automatic classification of sentiment in notes, where traditional machine learning algorithms were used. Another important area of suicide note research is the identification of forged suicide notes from genuine ones. Jones and Bennell (2007) have used supervised classification model and a set of linguistic features to distinguish genuine from forged suicide notes, achieving an accuracy of 82%.

Depression notes Work on identifying depression and other mental health conditions has become more prevelant over recent years, where a shared task was dedicated to distinguish depression and PTSD (Post Traumatic Stress Disorder) on Twitter using machine learning (Coppersmith et al., 2015). Morales et al. (2017) have argued that changes in cognition of people with depression can lead to different language usage, which manifests itself in the use of specific linguistic features. Research conducted byResnik et al. (2015) also used linguistic signals to detect depression with different topic modelling techniques. Work by Rude et al. (2004) used LIWC to analyse written documents by students who have experienced depression, currently depressed students as well as student who never have experienced depression, where it was found that individuals who have experienced depression used more first-person singular pronouns and negative emotion words. Nguyen et al. (2014) used LIWC to detect differences in language in online depression communities, where it was found that negative emotion words are good predictors of depressed text compared to control groups using a Lasso Model (Tibshirani, 1996). Research conducted by Morales and Levitan (2016) showed that using LIWC to identify sadness and fatigue helped to accurately classify depression.

Last statements Most work in the analysis of last statements of death row inmates has been conducted using data from The Texas Department of Criminal Justice, made available on their website (Texas Department of Criminal Justices, 2019). Recent work conducted by Foley and Kelly (2018) has primarily focused on the analysis of psychological factors, where it was found that specifically themes of *'love'* and *'spirituality'* were constant whilst requests for forgiveness declined over time. Kelly and Foley (2017) have identified that mental health conditions occur often in death row inmates with one of the most common conditions being depression. Research conducted by Heflick (2005) studied Texas last statements using qualitative methods and have found that often afterlife belief and claims on innocence are common themes in these notes. Eaton and Theuer (2009) studied qualitatively the level of apology and remorse in last statements, whilst also using logistic regression to predict the presence of apologies achieving an accuracy of 92.7%. Lester and Gunn III (2013) used the LIWC program to analyse last statements, where they have found nine main themes, including the affective and emotional processes. Also, Foley and Kelly (2018) found in a qualitative analysis that the most common themes in last statements were love (78%), spirituality (58%), regret (35%) and apology (35%).

3 Data

For our analysis and experiments we use three different datasets, which have been collected from different sources. For the experiments we use standard data preprocessing techniques and remove all identifying personal information.[1]

Last Statements Death Row This dataset has been made available by the Texas Department of Criminal Justices (2019) and contains 545 records of prisoners who have received the death penalty between 1982 and 2017 in Texas, U.S.A. A total of 431 prisoners wrote notes prior to their death.

[1]The authors are happy to share the datasets upon request

Due to the information available on this data we have done a basic analysis on the data available, hereafter referred to as ***LS***.

Suicide Note The data for this corpus has mainly been taken from Schoene and Dethlefs (2016), but has been further extended by using notes introduced by The Kernel (2013) and Tumbler (2013). There are total of 161 suicide notes in this corpus, hereafter referred to as ***GSN***.

Depression Notes We used the data collected by Schoene and Dethlefs (2016) of 142 notes written by people identifying themselves as depressed and lonely, hereafter referred to as ***DL***.

4 Linguistic Analysis

To gain more insight into the content of the datasets, we performed a linguistic analysis to show differences in structure and contents of notes. For the purpose of this study we used the Linguistic Inquiry and Word Count software (LIWC) (Tausczik and Pennebaker, 2010), which has been developed to analyse textual data for psychological meaning in words. We report the average of all results across each dataset.

Dimension Analysis Firstly, we looked at the word count and different dimensions of each dataset (see Table 1). It has previously been argued by Tausczik and Pennebaker (2010) that the words people use can give insight into the emotions, thoughts and motivations of a person, where LIWC dimensions correlate with emotions as well as social relationships. The number of *words per sentences* are highest in DL writers and lowest in last statement writers. Research by Osgood and Walker (1959) has suggested that people in stressful situations break their communication down into shorter units. This may indicate alleviated stress levels in individuals writing notes prior to receiving the death sentence. *Clout* stands for the social status or confidence expressed in a person's use of language (Pennebaker et al., 2014). This dimension is highest for people writing their last statements, whereas depressed people rank lowest on this. Cohan et al. (2018) have noted that this might be due to the fact that depressed individuals often have a lower socio-economic status. The *Tone* of a note refers to the emotional tone, including both positive and negative emotions, where numbers below 50 indicate a more negative emotional tone (Cohn et al., 2004). The tone for LS is highest overall and the lowest in DL, indicating a more overall negative tone in DL and positive tone in LS.

Type	GSN	LS	DL
Tokens per note	110.65	109.72	98.58
Word per Sent	14.87	11.42	16.88
Clout	47.73	67.68	19.94
Tone	54.83	75.43	25.51

Table 1: LIWC Dimension Analysis

Function Words and Content Words Next, we looked at selected function words and grammatical differences, which can be split into two categories called *Function Words* (see Table 2), reflecting how humans communicate and *Content words* (see Table 2), demonstrating what humans say (Tausczik and Pennebaker, 2010). Previous studies have found that whilst there is an overall lower amount of function words in a person's vocabulary, a person uses them more than 50% when communicating. Furthermore it was found that there is a difference in how human brains process function and content words (Miller, 1991). Previous research has found that function words have been connected with indicators of people's social and psychological worlds (Tausczik and Pennebaker, 2010), where it has been argued that the use of function words require basic skills. The highest amount of function words were used in DL notes, whilst both GSN and LS have a similar amount of function words. Rude et al. (2004) has found that high usage, specifically of first-person singular pronouns ("I") could indicate higher emotional and/or physical pain as the focus of their attention is towards themselves. Overall Just et al. (2017) has also identified a larger amount of personal pronouns in suicide-related social media content. Previous work by Hancock et al. (2007) has found that people use a higher amount of negations when also expressing negative emotions and used fewer words overall, compared to more positive emotions. This seem to be also true for the number of negations used in this case where amount of *Negations* were also highest in the DL corpus and lowest in the LS corpus, whilst the overall words count was lowest for DL and negative emotions highest. Furthermore, it was found that *Verbs*, *Adverb* and *Adjectives* are often used to communicate content, however previous studies have found (Jones and Bennell, 2007; Gregory,

1999) that individuals that commit suicide are under a higher drive and therefore would reference a higher amount of objects (through nouns) rather than using descriptive language such as adjectives and adverbs.

Type	GSN	LS	DL
Function	56.35	56.33	60.20
Personal pronouns	16.23	20.44	15.19
I	11.04	12.65	12.8
Negations	2.71	1.71	4.06
Verb	19.29	19.58	21.65
Adjective	4.45	2.58	4.98
Adverb	4.43	3.14	7.69

Table 2: LIWC Function and Content Words

Affect Analysis The analysis of emotions in suicide notes and last statements has often been addressed in research (Schoene and Dethlefs, 2018; Lester and Gunn III, 2013) The number of *Affect words* is highest in LS notes, whilst they are lowest in DL notes, this could be related to the emotional *Tone* of a note (see Table 1). This also applies to the amount of *Negative emotions* as they are highest in DL notes and *Positive emotions* as these are highest in LS notes. Previous research has analysed the amount of *Anger* and *Sadness* in GSN and DL notes and has shown that it is more prevalent in DL note writers as these are typical feelings expressed when people suffer from depression (Schoene and Dethlefs, 2016).

Type	GSN	LS	DL
Affect	9.1	11.58	8.44
Positive emotion	5.86	8.99	3.15
Negative emotion	3.15	2.58	5.21
Anger	0.61	0.65	1.03
Sadness	1.09	1.08	2.53

Table 3: LIWC Affect Analysis

Social and Psychological Processes *Social Processes* highlights the social relationships of note writers, where it can be seen in Table 4 that the highest amount of social processes can be found in LS and the lowest in DL. Furthermore LS notes tend to speak most about family relations and least about friends, this was also found by Kelly and Foley (2017) who found a low frequency in interpersonal relationships.

Type	GSN	LS	DL
Social processes	12.21	18.19	8.33
Family	1.17	2.17	0.47
Friends	0.77	0.38	0.73

Table 4: LIWC Social Processes

The term *Cognitive processes* encompasses a number of different aspects, where we have found that the highest amount of cognitive processes was in DL notes and the lowest in LS notes. Boals and Klein (2005) have found that people who use more cognitive mechanisms to cope with traumatic events such as break ups by using more causal words to organise and explain events and thoughts for themselves. Arguably this explains why there is a lower amount in LS notes as LS writers often have a long time to organise their thoughts, events and feelings whilst waiting for their sentence (Death Penalty Information Centre, 2019). *Insight* encompasses words such as *think* or *consider*, whilst *Cause* encompasses words that express reasoning or causation of events, e.g.: *because* or *hence*. These terms have previously been coined as *cognitive process words* by (Gregory, 1999), who argued that these words are less used in GSN notes as the writer has already finished the decision making process whilst other types of discourse would still try to justify and reason over events and choices. This can also be found in the analysis of our own data, where both GSN and LS notes show similar, but lower frequency of terms in those to categories compared to DL writers. *Tentativeness* refers to the language use that indicates a person is uncertain about a topic and uses a number of filler words. A person who use more tentative words, may have not expressed an event to another person and therefore has not processed an event yet and it has not been formed into a story (Tausczik and Pennebaker, 2010). The amount of tentative words used in DL notes is highest, whilst it is lowest in LS words. This might be due to the fact that LS writers already had to reiterate over certain events multiple times as they go through the process of prosecution.

Personal Concerns *Personal Concerns* refers to the topics most commonly brought up in the different notes, where we note that both *Money* and *Work* are most often referred to in GSN notes and lowest in LS notes. This might be due to the the fact that (Mind, 2013) lists these two topics as

Type	GSN	LS	DL
Cognitive Processes	12.19	10.85	16.77
Insight	2.37	2.3	4.07
Cause	0.95	0.8	1.94
Tentativeness	2.57	1.5	3.23

Table 5: LIWC Psychological Processes

some of the most common reasons for a person to commit suicide. *Religion* is most commonly referenced in LS notes, which confirms previous analysis of such notes (Foley and Kelly, 2018; Kelly and Foley, 2017) and lowest in DL notes. (Just et al., 2017) has found that the topic of *Death* is commonly referenced in suicide-related communication on Twitter. This was also found in this dataset, where GSN notes most commonly referenced death, whilst DL notes were least likely to reference this topic.

Type	GSN	LS	DL
Work	1.24	0.41	0.99
Money	0.68	0.18	0.31
Religion	0.82	2.7	0.09
Death	0.76	0.68	0.64

Table 6: LIWC Personal Concerns

Time Orientation and Relativity Looking at the *Time Orientation* of a note can give interesting insight into the temporal focus of attention and differences in verb tenses can show psychological distance or to which extend disclosed events have been processed (Tausczik and Pennebaker, 2010). Table 7 shows that the focus of LS letters is primarily in the past whilst GSN and DL letters focus on the present. The high focus on the past in DL notes as well as GSN notes could be, because these notes might draw on their past experiences to express the issues of their current situation or problems.The most frequent use of future tense is in LS letters which could be due to a LS notes writers common focus on afterlife (Heflick, 2005).

Type	GSN	LS	DL
Focus past	3.24	2.86	3.32
Focus present	14.39	1.43	16.11
Focus future	2.1	2.27	1.51

Table 7: LIWC Time orientation

Overall it was noted that for most analysis GSN falls between the two extremes of LS and DL.

5 Learning Model

The primary model is the Long-short-term memory (LSTM) given its suitability for language and time-series data (Hochreiter and Schmidhuber, 1997). We feed into the LSTM an input sequence $\mathbf{x} = (x_1, \ldots, x_N)$ of words in a document alongside a label $y \in Y$ denoting the class from any of the three datasets. The LSTM learns to map inputs x to outputs 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:

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

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 Graves (2013), this leads us to update our hidden state $\mathbf{h}$ as follows (where σ refers to the logistic sigmoid function and c is the 'cell state'):

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

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

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

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

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

A standard LSTM definition solves some of the problems of vanilla RNNs have (Hochreiter and

Schmidhuber, 1997), but it still has some shortcomings when learning long 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.

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 by Chang et al. (2017). This allows LSTMs to better learn input sequences and their dependencies and therefore temporal and complex data dependencies are learned on different layers. Whilst dilated LSTM alleviates the problem of learning long sequences, it does not contribute to identifying words in a sequence that are more important than others. Therefore we extend this network by (1) an embedding layer and (2) an attention mechanism to further improve the network's ability. A graph illustration of our learning model can be seen in Figure 2.

Dilated LSTM with Attention Each document D contains i sentences S_i, where w_i represents the words in each sentence. Firstly, we embed the words to vectors through an embedding matrix W_e, which is then used as input into the dilated LSTM.

The most important part of the dilated LSTM is the dilated recurrent skip connection, where $c_t^{(l)}$ is the cell in layer l at time t:

$$c_t^{(l)} = f(x_t^{(l)}, c_{t-s^{l-1}}^{(l)}). \tag{8}$$

$s^{(l)}$ is the skip length; or dilation of layer l; $x_t^{(l)}$ as the input to layer l at time t; and $f(\cdot)$ denotes a LSTM cell; M and L denote dilations at different layers:

$$s^{(l)} = M^{(l-1)}, l = 1, \dots L. \tag{9}$$

The dilated LSTM alleviates the problem of learning long sequences, however not every word in a sequence has the same meaning or importance.

Attention layer The attention mechanism was first introduced by Bahdanau et al. (2015), but has since been used in a number of different tasks including machine translation (Luong et al., 2015), sentence pairs detection (Yin et al., 2016) , neural image captioning (Xu et al., 2015) and action recognition (Sharma et al., 2015).

Our implemenetation of the attention mechanism is inspired by Yang et al. (2016), using attention to find words that are most important to the meaning of a sentence at document level. 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 sentence s is computed as follows, where u_{it} is the hidden representation of the dilated LSTM output, α_{it} represents normalised alpha weights measuring the importance of each word and S_i is the sentence vector:

$$u_{it} = \tanh(O + b_w) \tag{10}$$

$$\alpha_{it} = \frac{\exp\left(u_{it}^T u_w\right)}{\sum_t \exp\left(u_{it}^T u_w\right)} \tag{11}$$

$$s_i = \sum_t \alpha_{it} o\cdot \tag{12}$$

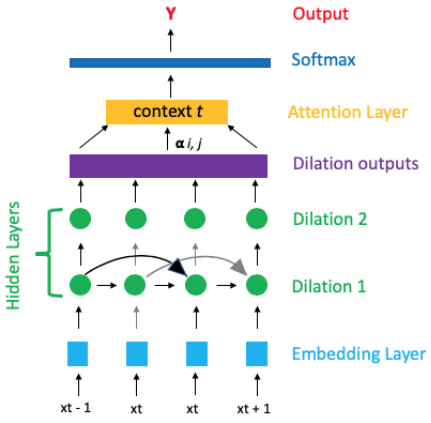

Figure 1: A 2-layer dilated LSTM with Attention.

6 Experiments and Results

For our experiments we use all three datasets, Table 8 shows the results for the experiments series. We establish three performance baselines on the datasets by using three different algorithms previously used on similar datasets. Firstly, we use the ZeroR and LMT (Logistic Model Tree) previously used by (Schoene and Dethlefs, 2016). Additionally we chose to benchmark our algorithm also against the originally proposed Bidirectional LSTM with attention proposed by Yang et al. (2016), which was also used on similar existing datasets before (Schoene and Dethlefs, 2018).

Furthermore we benchmark the Dilated Attention LSTM against two other types of recurrent neural networks. We use *200-dimensional* word embeddings as input into each network and all neural networks share the same hyper-parameters, where learning rate = 0.001, batch size = 128, dropout = 0.5, hidden size = 150 units and the *Adam* optimizer is used. For our proposed model - the Dilated LSTM with Attention - we establish the number of dilations empirically. There are 2 dilated layers with exponentially increasing dilations starting at *1*. Due to the size of the dataset we have split the data into *70%* training, *15%* validation and *15%* test data. We report results based on the test accuracy of the prediction results. It can be seen in Table 8 that the dilated LSTM with an attention layer outperforms the BiLSTM with Attention by 5.07%. Furthermore it was found that both the LMT and a vanilla bi-directional LSTM outperform a standard LSTM on this task. Previous results on similar tasks have yielded an accuracy of 69.41% using BiLSTM with Attention (Schoene and Dethlefs, 2018) and 86 % using a LMT (Schoene and Dethlefs, 2016).

7 Evaluation

In order to evaluate the DLSTM with attention we look in more detail at the predicted labels and visualise examples of each note to show which features are assigned the highest attention weights.

7.1 Label Evaluation

In Figure 2 we show the confusion matrix over the DLSTM with attention. It can be seen that LS notes are most often correctly predicted and DL notes are least likely to be accurately predicted.

The same applies to results of the main competing model (Bi-directional LSTM with Attention), Figure 3 shows that this model still misclassifies LS notes with DL notes.

7.2 Visualisation of attention weights

In order to see which features are most important to accurate classification we visualise examples from the test set of each dataset, where Figures 4, 5 and 6 show the visualisation of attention weights in the *GSN*, *LS* and *DL* datasets respectively. Furthermore, we also show three examples of the test data with typical errors the learning model makes in Figures 7 , 8 and 9. Words highlighted in darker shades have a higher attention weight.

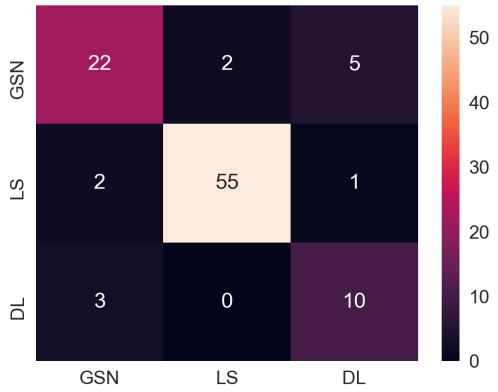

Figure 2: Confusion Matrix of test set labels - DLSTM Attention.

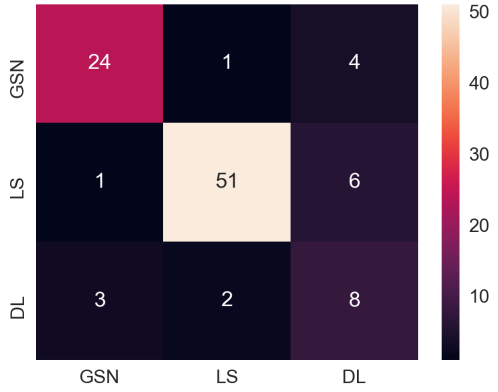

Figure 3: Confusion Matrix of test set labels - BiLSTM Attention.

The most important words highlighted in a last statement note (see Figure 4) are personal pronouns as well as an apology and expression of love towards friends and family members. This corresponds with the higher amount of personal pronouns, positive emotions and references to Family in LS notes compared to GSN and DL notes. Furthermore it can be seen that there is a low amount of cognitive process words and more action verbs such as *killing* or *hurt*, which could confirm that inmates have had more time to process events and thoughts and don't need cognitive words as a coping mechanism anymore (Boals and Klein, 2005).

Figure 5 shows a GSN note, where the most important words are also pronouns, references to family, requests for forgiveness and endearments. Previous research has shown that forgiveness is an important feature as well as the giving instructions such as *help* or phrases like *do not follow* are key

Model	Test Accuracy	Aver. Precision	Aver. Recall	Aver. F1-score
ZeroR	42.85	0.43	0.41	0.42
LMT	80.35	0.81	0.79	0.80
LSTM	62.16	0.63	0.61	0.62
BiLSTM	65.82	0.66	0.64	0.65
BiLSTM with Attention	82.27	0.85	0.83	0.84
DLSTMAttention	**87.34**	**0.88**	**0.87**	**0.87**

Table 8: Test accuracy and F1-score of different learning models in %

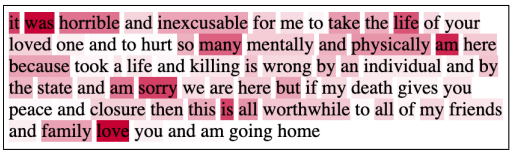

Figure 4: Example of LS note correctly classified.

to accurately classify suicide notes (Pestian et al., 2010). Terms of endearment for loved ones at the start or towards the end of a note (Gregory, 1999).

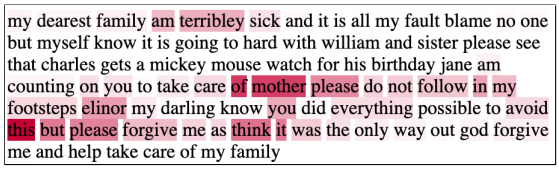

Figure 5: Example of GSN note correctly classified.

The DL note in Figure 6 shows that there is a greater amount of cognitive process verbs present, such as *feeling* or *know* as well as negations, which confirms previous analysis using LIWC.

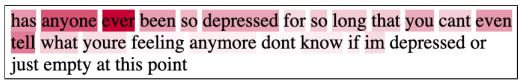

Figure 6: Example of DL note correctly classified.

Figure 7 shows a visualisation of a LS note. In this instances the word *God* was replaced with *up*, when looking into the usage of the word *up* in other LS notes, it was found that it was commonly used in reference to religious topics such as *God*, *heaven* or *up there*.

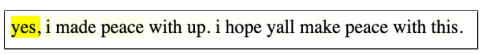

Figure 7: LS note error analysis

In Figure 8 a visualised GSN note is shown. Whilst there is still consistency in highlighting personal pronouns (e.g.: *you*), it can be seen that the end of the note is missing and more action verbs such as *hurt* or *take* are more important.

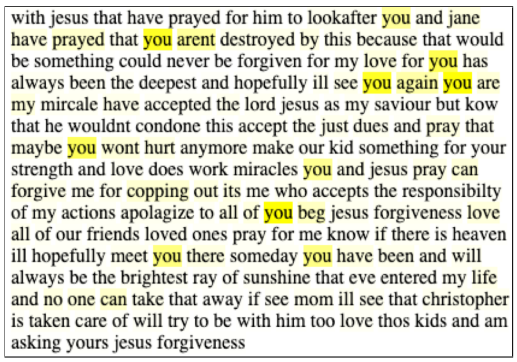

Figure 8: GSN note error analysis

The visualisation in Figure 9 demonstrates how the personal pronoun *I* has been removed from several DL notes, where DL notes are least likely to be predicted accurately as shown in Figure 2.

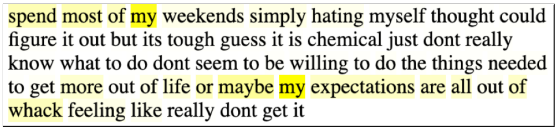

Figure 9: DL note error analysis

8 Conclusion

In this paper we have presented a new learning model for classifying long sequences. We have shown that the model outperforms the baseline by 6.99 % and by 5.07 % a competitor model. Furthermore we have provided an analysis of the linguistic features on three datasets, which we have later compared in a qualitative evaluation by visualising the attention weights on examples of each dataset. We have shown that the neural network pays attention to similar linguistic features as provided by LIWC and found in human evaluated related research.

References

D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proc. of the International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.

BBC. 2019. Facebook 'sorry' for distressing suicide posts on instagram.

Adriel Boals and Kitty Klein. 2005. Word use in emotional narratives about failed romantic relationships and subsequent mental health. *Journal of Language and Social Psychology*, 24(3):252–268.

Pete Burnap, Gualtiero Colombo, Rosie Amery, Andrei Hodorog, and Jonathan Scourfield. 2017. Multi-class machine classification of suicide-related communication on twitter. *Online social networks and media*, 2:32–44.

Rafael A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

Shiyu Chang, Yang Zhang, Wei Han, Mo Yu, Xiaoxiao Guo, Wei Tan, Xiaodong Cui, Michael Witbrock, Mark A Hasegawa-Johnson, and Thomas S Huang. 2017. Dilated recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 77–87.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. Smhd: A large-scale resource for exploring online language usage for multiple mental health conditions. *arXiv preprint arXiv:1806.05258*.

Michael A Cohn, Matthias R Mehl, and James W Pennebaker. 2004. Linguistic markers of psychological change surrounding september 11, 2001. *Psychological science*, 15(10):687–693.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39.

Malcolm Coulthard, Alison Johnson, and David Wright. 2016. *An introduction to forensic linguistics: Language in evidence*. Routledge.

Death Penalty Information Centre. 2019. Time on death row.

Judy Eaton and Anna Theuer. 2009. Apology and remorse in the last statements of death row prisoners. *Justice Quarterly*, 26(2):327–347.

Facebook. 2019. Suicide prevention.

SR Foley and BD Kelly. 2018. Forgiveness, spirituality and love: thematic analysis of last statements from death row, texas (2002-2017). *QJM: An International Journal of Medicine*.

Alex Graves. 2013. Generating Sequences With Recurrent Neural Networks. *CoRR*, abs/1308.0850.

Adam Gregory. 1999. The decision to die: The psychology of the suicide note. *Interviewing and deception*, pages 127–156.

Jeffrey T Hancock, Christopher Landrigan, and Courtney Silver. 2007. Expressing emotion in text-based communication. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 929–932. ACM.

Nathan A Heflick. 2005. Sentenced to die: Last statements and dying on death row. *Omega-Journal of Death and Dying*, 51(4):323–336.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Lstm can solve hard long time lag problems. In *Advances in neural information processing systems*, pages 473–479.

Natalie J Jones and Craig Bennell. 2007. The development and validation of statistical prediction rules for discriminating between genuine and simulated suicide notes. *Archives of Suicide Research*, 11(2):219–233.

Marcel Adam Just, Lisa Pan, Vladimir L Cherkassky, Dana L McMakin, Christine Cha, Matthew K Nock, and David Brent. 2017. Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nature human behaviour*, 1(12):911.

Brendan D Kelly and Sharon R Foley. 2017. Analysis of last statements prior to execution: methods, themes and future directions. *QJM: An International Journal of Medicine*, 111(1):3–6.

David Lester and John F Gunn III. 2013. Ethnic differences in the statements made by inmates about to be executed in texas. *Journal of Ethnicity in Criminal Justice*, 11(4):295–301.

Liu Yi Lin, Jaime E Sidani, Ariel Shensa, Ana Radovic, Elizabeth Miller, Jason B Colditz, Beth L Hoffman, Leila M Giles, and Brian A Primack. 2016. Association between social media use and depression among us young adults. *Depression and anxiety*, 33(4):323–331.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

David D Luxton, Jennifer D June, and Jonathan M Fairall. 2012. Social media and suicide: a public health perspective. *American journal of public health*, 102(S2):S195–S200.

George Armitage Miller. 1991. The science of words.

Mind. 2013. Depression.

Michelle Morales, Stefan Scherer, and Rivka Levitan. 2017. A cross-modal review of indicators for depression detection systems. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical PsychologyFrom Linguistic Signal to Clinical Reality*, pages 1–12.

Michelle Renee Morales and Rivka Levitan. 2016. Speech vs. text: A comparative analysis of features for depression detection systems. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 136–143. IEEE.

Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. 2014. Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing*, 5(3):217–226.

Bridianne O'dea, Mark E Larsen, Philip J Batterham, Alison L Calear, and Helen Christensen. 2017. A linguistic analysis of suicide-related twitter posts. *Crisis*.

Charles E Osgood and Evelyn G Walker. 1959. Motivation and language behavior: A content analysis of suicide notes. *The Journal of Abnormal and Social Psychology*, 59(1):58.

James W Pennebaker, Cindy K Chung, Joey Frazee, Gary M Lavergne, and David I Beaver. 2014. When small words foretell academic success: The case of college admissions essays. *PloS one*, 9(12):e115844.

John Pestian, Henry Nasrallah, Pawel Matykiewicz, Aurora Bennett, and Antoon Leenaars. 2010. Suicide note classification using natural language processing: A content analysis. *Biomedical informatics insights*, 3:BII–S4706.

John P Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, K Bretonnel Cohen, John Hurdle, and Christopher Brew. 2012. Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights*, 5:BII–S9042.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015. Beyond lda: exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 99–107.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

Annika M Schoene and Nina Dethlefs. 2018. Unsupervised suicide note classification.

Annika Marie Schoene and Nina Dethlefs. 2016. 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*, pages 128–133.

Andreas RT Schuck and Janelle Ward. 2008. Dealing with the inevitable: strategies of self-presentation and meaning construction in the final statements of inmates on texas death row. *Discourse & society*, 19(1):43–62.

Nabia Shahreen, Mahfuze Subhani, and Md Mahfuzur Rahman. 2018. Suicidal trend analysis of twitter using machine learning and neural network. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–5. IEEE.

Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. 2015. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Texas Department of Criminal Justices. 2019. Texas death row executions info and last words.

The Kernel. 2013. What suicide notes look like in the social media age.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tumbler. 2013. Suicide notes.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. 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*, pages 1480–1489.

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2016. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for Computational Linguistics*, 4:259–272.