Dilated LSTM with ranked units for Classification of Suicide Notes

Annika M Schoene
Department of Computer Science
The University of Hull
Kingston-upon-Hull, HU6 7RX
amschoene@gmail.com

Alexander P Turner
Department of Computer Science
The University of Hull
Kingston-upon-Hull, HU6 7RX
Alexander.Turner@hull.ac.uk

Nina Dethlefs
Department of Computer Science
The University of Hull
Kingston-upon-Hull, HU6 7RX
N.Dethlefs@hull.ac.uk

Abstract

Recent statistics in suicide prevention show that people are increasingly posting their last words online and with the unprecedented availability of textual data from social media platforms researchers have the opportunity to analyse such data. Furthermore, psychological studies have shown that our state of mind can manifest itself in the linguistic features we use to communicate. In this paper, we investigate whether it is possible to automatically identify suicide notes from other types of social media blogs in a document-level classification task. Also, we present a learning model for modelling long sequences, achieving an f1-score of *0.84* over the baselines of *0.53* and *0.80* (best competing model). Finally, we also show through visualisations which features the learning model identifies.

1 Introduction

The use of social media platforms, such as blogging websites has increasingly become part of everyday life and there is 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]. 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. These developments subsequently triggered the advances of new healthcare applications and methodologies that aid detection of concerning posts on social media platforms [Calvo et al., 2017].

In this paper we present a novel document-level classification task of suicide notes, depressed blog posts and 'neutral' blog posts. We introduce the dilated LSTM with ranked units for modelling long sequences, where previous research has found that the use of Dilated RNNs (DRNN) on sequence classification for language modelling [Chang et al., 2017] has outperformed competitive baselines such as standard LSTM/GRU architectures as well as more specialised models. We find in our experiment series that our learning model outperforms the baseline by 31 % and a competitor model by 4%. Whilst the analysis and classification of suicide notes and social media blogs has traditionally

AI for Social Good workshop at NeurIPS (2019), Vancouver, Canada.

been conducted separately, there have been studies investigating the differences of suicide and depression blogs [Schoene and Dethlefs, 2016, 2018]. This is mainly due to mental health conditions such as depression are often related to suicide [Mind, 2013]. To further extend this work we use a third corpus of social media blogs, which introduces a somewhat 'neutral' category where there is a range of different emotions and topics covered that may or may not occur in either suicide or depressed notes. We show through the visualisation of attention weights for each document type which features are most indicative for accurate classification. Arguably, the exploration and comparison of suicide notes with depressed and 'neutral' blogs, could help us to find further differentiating factors and aid in identifying suicidal ideation in textual data.

2 Related Work

The World Health Organisation [WHO, 2019] outline in a recent report that suicide is the second leading cause of death for people aged 15-29 worldwide and reducing the rate of suicide worldwide is listed on the Sustainable Development Goals. Furthermore, there has been a trend recognised that especially young people tend to publish their suicide notes or express their suicidal feelings online [Desmet and Hoste, 2013]. Work on suicide notes has often focused on identifying suicidal ideation online [O'dea et al., 2017, Shahreen et al., 2018] or distinguishing genuine from forged suicide notes [Coulthard et al., 2016]. Similarly, work on identifying mental health conditions online has focused on using Twitter to distinguish depression and PTSD (Post Traumatic Stress Disorder) [Coppersmith et al., 2015]. Whilst, work on classifying blogs from social media platforms has focused on predicting sentiment or emotions [Binali et al., 2010] or characteristics of the author of a blog, such as age [Rosenthal and McKeown, 2011] or gender [Bartle and Zheng, 2015]. Previous work in suicide note classification has focused both on traditional machine learning to distinguish suicide notes from depressed and love notes Schoene and Dethlefs [2016]. Furthermore, research by Schoene and Dethlefs [2018] used a Bidirectional LSTM with attention to classify suicide notes, using emotion sequences and achieving an accuracy of 75.29%.

Recurrent neural networks (RNNs) are well suited towards natural language processing tasks due to their ability to handle sequential data [Hochreiter and Schmidhuber, 1997], however there are still shortcomings which ultimately effect the accurate classification of longer sequences. This is mainly due to the problem of *vanishing* and *exploding gradient descent* [Bengio et al., 1994], which impacts on the RNNs ability to maintain mid and short term memory when memorising long-term dependencies. Various approaches have tried to solve the problem of learning long-term dependencies in temporal data, where variations of multiscale RNNs have produced state-of-the-art results on various tasks. Generally speaking multiscale RNNs, group the hidden units of the network into multiple modules that operate on different timescales [El Hihi and Bengio, 1996, Koutnik et al., 2014, Chung et al., 2016] in order to overcome this problem.

3 Methods

3.1 Data and Preprocessing

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]

The genuine suicide notes have been mainly taken from Schoene and Dethlefs [2016], but has been further extended by using notes made publicly available online Tumbler [2013], The Kernel [2013]. There are total of 211 suicide notes in this corpus, hereafter referred to as ***GSN***. For the depressed posts, we used the data collected by Pirina and Çöltekin [2018] and randomly selected 211 posts (referred to as ***DL***). Similarly, for 'neutral' blog posts we used data made available by Schler et al. [2006] and randomly selected 211 blog posts (referred to as ***BL***). Due to the variable length of both depressed and 'neutral' blogs, we have chosen to exclude any documents that surpass the overall length of 1500 words in order to preserve similarities to GSN notes. Previous work in this area has predominantly focused on distinguishing suicide notes from other types of notes that are in a distinct category, e.g.: depression or love notes. However, when attempting to classify suicides notes, they

[1]The authors are happy to share the datasets used in this task upon request.

usually do not come in neat types of categories and therefore we have chosen a random sample of blog posts to make the task more applicable to real-world scenarios. Furthermore, classifying suicide notes in such a setup could lead to identify further distinguishing features in the language used in these notes.

3.2 Learning Algorithm

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.

Dilated LSTM with ranked units 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 $LSTM_t^{(l)}$ is the cell in layer l at time t:

$$LSTM_t^{(l)} = f(x_t^{(l)}, c_{t-s^{l-1}}^{(l)}) \cdot \tag{1}$$

$s^{(l)}$ is the skip length; or dilation of layer l; $x_t^{(l)}$ as the input to layer l at time t; M and L denote dilations at different layers: $s^{(l)} = M^{(l-1)}, l = 1, \dots L$.

The dilated LSTM alleviates the problem of learning long sequences, but not each document has the same sequence length, so in order to overcome this variability we provide fixed boundaries to each layer by reducing the number of hidden units per sub-LSTM hierarchically. Therefore larger sub-LSTMs focus on learning long-term dependencies, whilst smaller sub-LSTMs focus on more frequently occurring short-term dependencies. This leads to improved performance as it has been shown in other contexts [El Hihi and Bengio, 1996, Chung et al., 2016].

We extended the earlier implementation with an attention mechanism 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:

$$u_{it} = \tanh(O + b_w) \tag{2}$$

$$\alpha_{it} = \frac{\exp\left(u_{it}^T u_w\right)}{\sum_t \exp\left(u_{it}^T u_w\right)} \tag{3}$$

$$s_i = \sum_t \alpha_{it} o \cdot \tag{4}$$

4 Experiments

For our experiments we use all three datasets, where we establish two performance baselines on the datasets. Firstly we use a Maximum Entropy classifier due to its suitability to textual data where conditional independence of the features can't be assumed. Additionally we chose to benchmark our algorithm against the originally proposed Bidirectional LSTM with attention proposed by Yang et al. [2016], as it also utilises attention. Furthermore we benchmark the Dilated LSTM with ranked units against two other types of RNNs.

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 = 64, dropout= 0.5 and the *Adam* optimiser is used. Furthermore we use the full sequence length of each document as input.

For our proposed model - the Dilated LSTM with ranked units - we establish the number of dilations empirically. There are 2 dilated layers with exponentially increasing dilations starting at *1*. The number of hidden units is adjusted according to the sequence length used as input to each sub-LSTM, where the number of hidden units is always half of the given sequence length. For example, given a sequence length of 160 and 2 dilations the input length to the sub-LSTM is [160,80], whilst the number of hidden units adjusts from 80 to 40. For all other learning models the number of hidden units is set to 300.

Due to the size of the dataset we split the data into *70%* training, *15%* validation and *15%* test data.

4.1 Results

All results are shown in Table 1 and we use precision, recall and f1-score as our evaluation metrics. It can be seen that the Dilated LSTM with ranked units and an attention layer outperforms both established benchmarks by 31% (Maximum Entropy) and 4% (BiLSTM with attention) respectively. Of particular interest are the results of the vanilla LSTM as they are considerably below the Maximum Entropy classifiers baseline and the next related model, the Bidirectional LSTM. Taking into account earlier observations that LSTMs may struggle to learn sequences above a certain length, we conducted an additional experiment where the sequence length was restricted to 300. This experiment yielded substantially better results with an f1-score of *0.66*. However, this has also meant that over 50% of the documents used in these experiments were cut short and not all information available was utilised.

Table 1: Experiment results of different learning models using precision, recall and f1-score

Learning Model	Aver. Precision	Aver. Recall	Aver. F1-score
Maximum Entropy	0.763	0.60	0.53
LSTM	0.50	0.46	0.31
BiLSTM	0.83	0.76	0.74
BiLSTM with attention	0.84	0.81	0.80
DLSTMattention	0.82	0.81	0.81
DilatedLSTM ranked units	**0.89**	**0.84**	**0.84**

Evaluation In order to see which features are most important to accurate classification we visualise attention weights and show examples from the test set of each dataset (see Figures 1, 2 and 3 in Appendix A), where words highlighted in darker shades have higher attention weights. One of the main differences in these three types of documents it the usage of personal pronouns, where in GSN notes there is frequent usage of *'you'*, whilst both other documents mainly refer to the first person singular or plural. Furthermore it can be seen that there is a range of different topics and emotions present in each document. More specifically, it can be seen that in GSN notes emotions such as *love*, *joy* and *peacefulness* are present, whilst in DL blogs *anger* and *hate* are predominant. Furthermore it can be seen that in the DL blog suicidal ideation is mentioned, however from a linguistic and sentiment perspective it is clearly distinct from a GSN note.

5 Conclusion

In this paper we have introduced the Dilated LSTM with ranked units and shown that the learning model is able to successfully distinguish suicide notes from both depressed blogs and 'neutral' blogs. Therefore demonstrating that accurate classification is possible solely on linguistic patterns in this type of textual data. Furthermore, we have shown by visualising attention weights which words are most important to each text category. However, additional research is needed to understand if, for example, these language patterns generalise over larger datasets and which role emotions expressed in textual data could help further to identify suicidal ideation. Given further research is conducted such work could be useful in a number of scenarios, including but not limited to assessing the seriousness of a suicide attempt in a clinical setting, distinguishing forged from genuine notes or help suicide prevention charities in flagging up emails that indicate high risk of suicide. Finally, we are currently testing how the learning model performs on other types of textual social media data and tasks such as sentiment classification of tweets.

References

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

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. Association between social media use and depression among us young adults. *Depression and anxiety*, 33(4):323–331, 2016.

Facebook. Suicide prevention, 2019. URL `https://www.facebook.com/help/594991777257121/`.

BBC. Facebook 'sorry' for distressing suicide posts on instagram, 2019. URL `https://www.bbc.co.uk/news/uk-46976753`.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 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, 2015.

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

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

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

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

Annika Marie Schoene and Nina Dethlefs. 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, 2016.

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

Mind. Depression, 2013. URL `https://www.mind.org.uk/information-support/types-of-mental-health-problems/depression/#.XNFba5NKiqQ`.

WHO. Sustainable development goal 3, 2019. URL `https://sustainabledevelopment.un.org/sdg3`.

Bart Desmet and VéRonique Hoste. Emotion detection in suicide notes. *Expert Systems with Applications*, 40(16):6351–6358, 2013.

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

Nabia Shahreen, Mahfuze Subhani, and Md Mahfuzur Rahman. 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, 2018.

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

Haji Binali, Chen Wu, and Vidyasagar Potdar. Computational approaches for emotion detection in text. In *4th IEEE International Conference on Digital Ecosystems and Technologies*, pages 172–177. IEEE, 2010.

Sara Rosenthal and Kathleen McKeown. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 763–772. Association for Computational Linguistics, 2011.

Aric Bartle and Jim Zheng. Gender classification with deep learning. In *Technical report*. The Stanford NLP Group., 2015.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Yoshua Bengio, Patrice Simard, Paolo Frasconi, et al. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

Salah El Hihi and Yoshua Bengio. Hierarchical recurrent neural networks for long-term dependencies. In *Advances in neural information processing systems*, pages 493–499, 1996.

Jan Koutnik, Klaus Greff, Faustino Gomez, and Juergen Schmidhuber. A clockwork rnn. *arXiv preprint arXiv:1402.3511*, 2014.

Junyoung Chung, Sungjin Ahn, and Yoshua Bengio. Hierarchical multiscale recurrent neural networks. *arXiv preprint arXiv:1609.01704*, 2016.

Tumbler. Suicide notes, 2013. URL `http://suicide--notes.tumblr.com/`.

The Kernel. What suicide notes look like in the social media age, 2013. URL `https://kernelmag.dailydot.com/features/report/6451/what-suicide-notes-look-like-in-the-social-media-age/`.

Inna Pirina and Çağrı Çöltekin. Identifying depression on reddit: The effect of training data. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 9–12, 2018.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205, 2006.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 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, 2016.

A Appendix

farewell letter no more joy no more joy no more love no more sun or moon to see a little bit nasty just a corpse not very nice for you either reckon :UNK: sun gives warmth love strength :UNK: moon is cold and white clouds deprive :UNK: sun of its strength but :UNK: night is clear and bright ive often dreamt of beautiful things all ive found is smiles if ever rebelledi gained nothing just pain and anguish it sounds resigned thats what am life has stolen my life awayit can all be so simple but went off course built up stupid hopes such a pity about my lovelove was string and beautifulbut time is strongerit makes you forget may be forgiven for my fit of sentimentality wouldnt have made a very good poet

Figure 1: Examples of correctly classified 'GSN' note

dont know if doing this right but putting it down on paper gets it out of my head at least temporarily and im dying inside im trapped feel so low unwelcome familiar thoughts in my head and nowhere to turn my two teenage kids in bed my bloke away at his kids for weekend want to cry but cant anymore dont want to bring anyone down with my feelings but so lonely at moment life has nothing new to offer me and it seems so easy to leave it all but how can when have two beautiful kids upstairs cant do that to them and thats why feel so low hate life always have feel comfortable in depression after a while and thats when get to point of not caring who hurt had several episodes through teens and adult life drink used to help unemployment let me be selfish being single mom meant didnt have to explain my feelings but this is first episode since been with partner of 3 years he wont understand and just dread tomorrow when he comes home just want to sleep and never wake up trapped angry im trapped feel selfish for feeling this way and that makes me even angrier any parents out there understand what mean

Figure 2: Examples of correctly classified 'DL' note

today started two songs in garage band for our portfolio that we are creating like both of them and thats bout it cya got to go home now jo is next to me today and im writing bout me so there but havnt got much to talk bout her really hmmm jo is 1 of :UNK: grossest people know in :UNK: world and have known her for a very very very long time like 11 years or something1 like wowo same as ang really know her for that long to crazy stuff goin to see john buttler tonight hello and yes im in computing class writing my diary for :UNK: whole world to see smart isnt it p well what an tell you all other than am in computing well am in a band at :UNK: moment we are called shye but we dont like that name any moer well jim nad amy dont emma and dont really care as long as we are playing so if any 1 has a really cool band name for a light rock group of four girls let me know

Figure 3: Examples of correctly classified 'neutral' blog