Hierarchical Multiscale Recurrent Neural Networks for Detecting Suicide Notes

Article *in* IEEE Transactions on Affective Computing · February 2021

DOI: 10.1109/TAFFC.2021.3057105

Hierarchical Multiscale Recurrent Neural Networks for Detecting Suicide Notes

Annika M Schoene, Alexander P Turner, Geeth De Mel and Nina Dethlefs

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 two document-level classification tasks. The first task aims to identify suicide notes from depressed and blog posts in a balanced dataset, whilst the second experiment looks at how well suicide notes can be classified when there is a vast amount of neutral text data, which makes the task more applicable to real-world scenarios. Furthermore we perform a linguistic analysis using LIWC (Linguistic Inquiry and Word Count). We present a learning model for modelling long sequences in two experiment series. We achieve an f1-score of *88.26%* over the baselines of *0.60* in experiment 1 and *96.1%* over the baseline in experiment 2. Finally, we show through visualisations which features the learning model identifies, these include emotions such as love and personal pronouns.

Index Terms—Natural Language Processing, Recurrent Neural Networks, Text Classification

1 Introduction

Whilst both machine and deep learning techniques have been predominantly used for commercial purposes, there has also been an increased awareness of how AI approaches could contribute to solving some of the biggest social problems humans face worldwide [1]. This awareness has led to the creation of new workshops and conferences that fall under the umbrella of *AI for Social Good*, where machine learning researchers connect with Non-Governmental Organisations (NGOs), charities and other problem owners to create practical solutions. These problems and challenges are usually closely linked to accelerating progress towards the UN Sustainable Development Goals (SGD) produced by the United Nations (UN) [2]. These goals include, but are not limited to protecting democracy, education, social welfare and justice as well as health care and environmental sustainability.

Especially within the SGD for health care, there is an increased focus on mental health. In a recent report, the World Health Organisation [2] outlines that suicide is the second leading cause of death for people aged 15-29 worldwide. Reducing the rate of suicide worldwide has therefore been listed as one of the objectives of the Sustainable Development Goals for health care. It is estimated that around 25-30% of people who died by suicide leave behind a suicide note, however, this figure can be as high as 50% depending on cultural or ethnic differences in demographics [3]. [4] have found that there is an increasing trend amongst younger people to publish their suicide notes or express their suicidal feelings online. Furthermore, psychological studies have shown that our state of mind can manifest itself in the linguistic features we use to communicate [5], [6]. At the same time, 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 and other mental health conditions. Whilst there are efforts to tackle suicide and other mental health conditions online by social media platforms such as Facebook [7], there are still concerns that there is not enough support and protection, especially for younger users [8].

Taking these trends into account and with this unprecedented availability of textual data from social media platforms researchers have now the opportunity to analyse such data and use their findings in several different application areas. This has led to a notable increase in research of suicidal and depressed language usage [9], [10] and subsequently triggered the development of new healthcare applications and methodologies that aid detection of concerning posts on social media platforms [11]. Traditionally, work on suicide notes has focused on distinguishing genuine from forged suicide notes in the field of forensic linguistics, where the findings were used as additional evidence in legal proceedings [12]. However, in recent years and with the advances in machine and deep learning, there has been an increasing amount of research conducted to identify suicidal ideation or suicide notes in online settings, such as social media platforms [13], [14].

In this paper, there will firstly be an exploration of existing research and literature in the field of suicide note detection in section 2. Then there will be an analysis of the linguistic features for the different datasets used in section 3. In section 4 we will introduce the learning model, a dilated LSTM with attention. Next, there will be a series of two experiments using two different kinds of datasets and a variety of recurrent neural networks in section 5. For the first experiment series we use a balanced dataset to classify suicide notes, depressed posts and blog posts to see how hard the task proves in this setting. The second experiment aims to make the task more applicable to the real world and both depressed and blog posts are increased to reflect the rarity of genuine suicide notes on social media platforms. In section 6 we discuss the experimental results and evaluate the visualisations.

2 Related Work

Over the years there has been much research conducted into the accurate classification of suicide notes or detection of suicidal ideation online [13], [14], where researchers use several different methodologies including but not limited to traditional machine learning [15], deep learning [16] and sentiment analysis [17]. Such research has been conducted in a range of different disciplines like psychology [18], linguistics [13] or healthcare [19]. Many experiments have also been conducted comparing different types of textual data with suicide notes such as depressed language or blog posts [20]. Overall there has been a growing interest in looking at content created online that may solicit need for help [21] or detecting mental health issues [22]. This literature review will focus on introducing work looking at the classification of suicide notes and suicidal ideation detection, but also review work in the space of depressed language and last statements due to the nature of the experiments.

2.1 Suicide note classification

The analysis of suicide notes has been used in various academic settings such as psychology or forensic linguistics to either identify the genuineness of a suicide note or to predict the state of mind of a note writer [12]. It has been argued in previous research that our drive or motivation affects how we communicate and therefore it is believed that our spoken and written language represents those shifting psychological states [23]. This argument has been taken further by [6] who suggested that there is a shift in one's linguistic expression due to the aroused cognitive state suicidal individuals experience. These findings have led to [4]'s argument that there is an increased need for 'automatic procedures that can spot suicidal messages and allow stakeholders to quickly react to online suicidal behaviour or incitement'. Therefore recent research has looked at different aspects of suicide notes to find out what "makes" a suicide note, where identifying linguistic features and patterns, affective states or specific emotions as well as dominant topics have been used in different analyses and experiments. [24] provide an overview of applications, methods and domains in suicide note research.

One of the settings in which the validation of a suicide note is important is in court cases or hearings where expert evidence is given by professionals such as forensic linguists to verify the author of the note or its genuineness [12]. Another field where the analysis of suicide notes is crucial is psychology, where one of the most commonly cited studies has been conducted by [25]. In their study, they collected a corpus of 33 genuine suicide notes and another set of 33 suicide notes that were forged. Their analysis showed that there was a clear difference in language used, which made the genuine notes distinctive when compared to the forged notes. This study has been used as a foundation for many other studies afterwards [26] and researchers such as [5] have compared this set of suicide notes with a set of normal letters to friends. Whilst especially early work in linguistics and psychology has mainly focused on the distinguishing factors of linguistics and topics [27], the availability of such data to researchers from other disciplines has opened up opportunities to use traditional machine learning and feature engineering for classifying suicide notes.

[28] have used a supervised classification model and a set of linguistic features to distinguish genuine from forged suicide notes, achieving an accuracy of 82%. Studies using traditional machine learning have been taken further recently by [29] who also used a set of suicide notes and correctly hypothesised that when applying the set to a machine learning algorithm it would outperform mental health professionals in classifying suicide notes correctly. Detecting affective states or emotions in such data has also grown in popularity. Particularly the work of [10] has been influential in the field and in their study they have found that there are fifteen different emotional concepts which prove to be significant in identifying genuine suicide notes. These fifteen sentiment features have also been used by [30] in the i2b2/VA/ Cincinnati Medical Natural Language Processing Challenge. The challenge aimed to develop a model which could automatically identify emotions on sentence-level of a suicide note. The hybrid model developed by [30]) achieved an accuracy of 61.39% in detecting emotions using various techniques such as machine learning-based emotion classification. [30] argue that one of the key factors for successful identification of emotions is to split the 15 pre-specified emotions into three different classes (positive, negative and neutral). [31] have focused on combining both sentiment and linguistic features which led to achieving a test accuracy of 86.6%. [32] have used four different feature groups including sentiment to assess suicide risk using a hybrid model.

Suicide note research has not only focused on the sentiment conveyed in notes but also on linguistic [5] and content [33] features. Research conducted by [28] used Receiver Operating Characteristic (ROC) Analysis to distinguishing genuine and forged suicide notes from each other, yielding an average accuracy of 0.82 AUC. Other work conducted by [31] has found that using a combination of both linguistic and sentiment features achieves an accuracy of 86.61% by using a logistic model tree (LMT).

2.2 Suicide ideation classification

Recent years have seen an increase in the analysis of suicidal ideation on social media platforms, such as Twitter. [14] 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 [34] has developed a classifier to distinguish suicide-related themes such the reports of suicides and casual references to suicide. The increased use of deep learning in other areas of Natural Language Processing [35] has also led to more studies using Recurrent Neural Networks (RNN) or Convolutional Neural Networks (CNN) to detect suicide notes or suicidal ideation [36]. Work by [37] used multiple neural network architectures to detect suicidal ideation. Research by [38] uses multi-task learning to estimate the risk of suicide using multiple public datasets from various shared tasks. Work by [39] has looked at identifying suicidal ideation on Twitter by using lexical, structural and sentiment features, using traditional machine learning and achieved an F-measure of 0.728.

2.3 Depression notes

Work on identifying depression and other mental health conditions has become more prevalent over recent years, where a shared task was dedicated to distinguishing depression and Post Traumatic Stress Disorder (PTSD) on Twitter using machine learning [9]. [40] have argued that changes in the cognition of people with depression can lead to different language usage, which manifests itself in the use of specific linguistic features. Research conducted by [41] also used linguistic signals to detect depression with different topic modelling techniques. Work by [42] used the Linguistic Inquiry and Word Count software (LIWC) to analyse written documents by students who have experienced depression, currently depressed students as well as students 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. [43] 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 [44]. Research conducted by [45] showed that using LIWC to identify sadness and fatigue helped to accurately classify depression. [46] use Convolutional Neural Networks to model the relationship between depression and people who attempt suicide. Some work has focused on detecting mental health signals related to other conditions such as bipolar disorder, major depressive disorder, post-traumatic stress disorder and seasonal affective disorder [47]. In their work [48] have looked extensively at the which features are relevant when classifying depression in tweets.

2.4 Social Media blogs

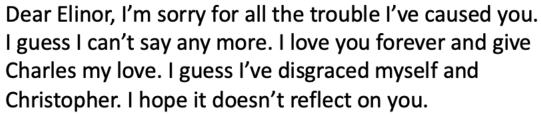
Work on classifying blogs from social media platforms has focused on predicting sentiment or emotions [49] or characteristics of the author of a blog, such as age [50] or gender [51]. Other work has focused on modelling ideologies in blogs using topic modelling techniques [52].

3 Data

This section provides an overview of the different datasets used as well as where and how they have been collected. All corpora have been anonymised in order to protect the authors' identity and those mentioned in their communication, which includes any places, names or references to identifying information. The examples of notes below have been chosen for their brevity, many of the notes in the corpus are of greater length[1]. 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 [31]. However, when attempting to classify suicides notes from 'neutral' blog posts is harder, because they usually do not come in neat types of categories. 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 help to identify further distinguishing features in the language used in these notes. Below we outline the different datasets used in the subsequent experiments and examples of the notes can be seen in Figures 1, 2, 3:

3.1 Genuine Suicide Note Data

Genuine suicide notes provide a unique insight into the mindset of a person who has died by suicide [53]. Therefore we have chosen to only use genuine suicide notes in our experiments and made a conscious decision not to use other datasets such as Twitter suicide datasets [39]. The main reason for this being that these tweets have mainly been collected using specific keywords such as *'suicide'* to accumulate the data and there is no human verification that the person who wrote this tweet is indeed suicidal or has passed away. Due to the sparsity of genuine suicide notes that are publicly available, we have added new genuine suicide notes to the corpus provided by [20]. Other new additions to this corpus includes data from various sources (for a full list, see Appendix A). There is a total of 211 genuine suicide notes (hereafter **GSN**, see Figure 1) used in these experiments.

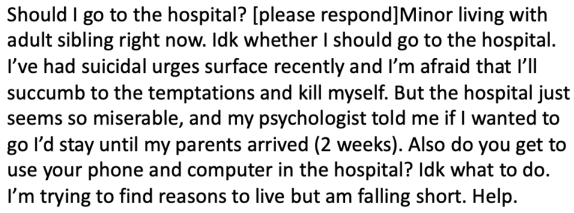


Dear Elinor, I'm sorry for all the trouble I've caused you. I guess I can't say any more. I love you forever and give Charles my love. I guess I've disgraced myself and Christopher. I hope it doesn't reflect on you.

Figure 1: Example of a suicide note.

3.2 Depression Notes

We used the Reddit depression data provided by [54] to create two different datasets for the two experiments. The first dataset consists of 211 depressed notes, hereafter referred to as **DL1** and the second dataset includes 1293 depressed notes (hereafter referred to as **DL2**, see Figure 2).



Should I go to the hospital? [please respond]Minor living with adult sibling right now. Idk whether I should go to the hospital. I've had suicidal urges surface recently and I'm afraid that I'll succumb to the temptations and kill myself. But the hospital just seems so miserable, and my psychologist told me if I wanted to go I'd stay until my parents arrived (2 weeks). Also do you get to use your phone and computer in the hospital? Idk what to do. I'm trying to find reasons to live but am falling short. Help.

Figure 2: Example of a depressed post.

3.3 Neutral Blog Posts

We have chosen a random number of online blog posts as our neutral category, which were collected by [55]. For both types of experiments we used 211 blog posts (hereafter referred to as **NEU1**) and 3500 examples of blog posts (hereafter referred to as **NEU2**). We have chosen this amount of blog posts empirically to ensure that the overall amount

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

of GSN notes is below 5% to make the task more applicable to the real world, an example of a neutral blog post can be seen in Figure 3.

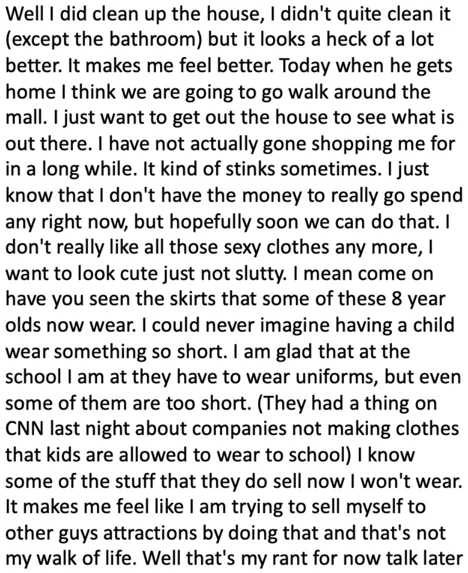

Well I did clean up the house, I didn't quite clean it (except the bathroom) but it looks a heck of a lot better. It makes me feel better. Today when he gets home I think we are going to go walk around the mall. I just want to get out the house to see what is out there. I have not actually gone shopping me for in a long while. It kind of stinks sometimes. I just know that I don't have the money to really go spend any right now, but hopefully soon we can do that. I don't really like all those sexy clothes any more, I want to look cute just not slutty. I mean come on have you seen the skirts that some of these 8 year olds now wear. I could never imagine having a child wear something so short. I am glad that at the school I am at they have to wear uniforms, but even some of them are too short. (They had a thing on CNN last night about companies not making clothes that kids are allowed to wear to school) I know some of the stuff that they do sell now I won't wear. It makes me feel like I am trying to sell myself to other guys attractions by doing that and that's not my walk of life. Well that's my rant for now talk later

Figure 3: Example of a blog post.

3.4 Linguistic Analysis

To gain more insight into the content of the datasets, we performed a linguistic analysis to show differences in the structure and contents of the datasets. For this study, we used the Linguistic Inquiry and Word Count software (LIWC) [56], which has been developed to analyse textual data for psychological meaning in words. We report the average of all results across each dataset. LIWC has been used in previous research to annotate datasets for suicide risks in addition to experts to determine linguistic profiles of suicide-related Twitter posts [57]. Other work by [42] used LIWC to analyse written documents by students who have experienced depression, currently depressed students as well as students who never have experienced depression.

3.4.1 Dimension Analysis

Firstly, we looked at the word count and different dimensions of each dataset (see Table 1).

Type	GSN	NEU1	NEU2	DL1	DL2
Word Count	155.43	198.89	247.70	180.25	182.78
Word per Sent	16.20	20.34	18.32	17.32	17.83
SixItr	12.10	16.48	17.09	13.99	13.91
Analytic	32.85	53.19	50.95	29.04	25.91
Clout	46.73	45.08	47.54	23.64	22.88
Authentic	64.21	55.93	54.51	82.18	81.65
Tone	54.67	53.09	51.60	23.25	23.06

Table 1: LIWC Dimension Analysis

It has previously been argued by [56] that the words people use can give insight into the emotions, thoughts and motivations of a person, where LIWC dimensions correlate emotions as well as social relationships. The number of *words per sentences* are highest in DL writers and lowest in GSN notes. Research by [5] has suggested that people in stressful situations break their communication down into shorter units. This may indicate alleviated stress levels in individuals writing notes before taking their own life. *Clout* stands for the social status or confidence expressed in a person's use of language [58]. This dimension is highest for people writing blog posts, whereas depressed people rank lowest on this. [59] have noted that this might be because depressed individuals often have a lower socioeconomic 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 [60]. The tone for GSN is highest overall and the lowest in DL, indicating a more overall negative tone in DL and positive tone in GSN. It was also found by [57] that the most concerning suicide-related content found on Twitter also had a higher word count, here the highest word count is in blog posts, whilst GSN notes have the lowest word count. This could be due to the fact that there is no character restriction placed upon bloggers. *SixItr* in Table 1 refers to words that are longer than 6 letters and are meant to indicate the social class and level of education of a person. It can be seen that the lowest scores were observed by GSN and the highest by blog posts writers. Whilst there is no additional information available to evaluate both educational or socio-economic factors, it could be argued that the lack of longer words in GSN notes is due to the argument made by [5] that GSN writers break communication down to shorter units due to the alleviated stress-levels and not their educational or socio-economic background. The *Analytical thinking dimension* indicates to which extent people use "formal, logical, and hierarchical thinking patterns" [56]. NEU writers score highest in this category and GSN writers score lowest. It has been found that people who score low in analytical thinking tend to write and use spoken language more narratively and focus on the present as well as personal experiences, compared to people who score highly in this [58]. The term *Authenticity* refers to which extent people write about themselves in an honest way, where they are typically portrayed as more humble, personal and vulnerable [61]. DL notes were the most authentically written whilst the least authentic words were written by NEU writers. Arguably blog posts do not require a writer to be vulnerable and with the increasing amount of blogging as a marketing tool there may be less personal or humble language found in these posts.

3.4.2 Function Words and Content Words

The next section looks 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 [56].

Type	GSN	NEU1	NEU2	DL1	DL2
Function	56.80	47.87	49.12	58.27	59.35
Personal pronouns	15.85	10.06	10.23	14.35	14.32
I	10.64	6.45	6.26	11.63	11.60
Negations	2.87	1.47	1.65	3.10	3.24
Verb	19.06	16.46	15.92	21.10	21.40
Adjective	4.54	4.71	4.25	4.80	4.82
Adverb	4.79	5.27	5.64	6.91	7.20

Table 2: LIWC Function and Content Words

Function words refer to a variety of different word categories, such as pronouns or auxiliary verbs and make up the majority of all words that are persons uses [56]. It was found that there is a difference in how human brains process function and content words [62]. Research has also found that function words have been connected with indicators of people's social and psychological worlds [56], 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 blog posts have the least amount of function words. [42] 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 [57] has also identified a larger amount of personal pronouns in suicide-related social media content. This may be the reason why GSN notes are high in personal pronouns overall and the first-person singular, whilst blog posts are lowest in both categories. Previous work by [63] 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 seems to be also true for the number of negations used in this case where the number of *Negations* were also highest in the DL corpus and lowest in the blog corpus, whilst negative emotions are also highest in DL notes. Furthermore, it was found that *Verbs*, *Adverb* and *Adjectives* are often used to communicate content, however previous studies have found [28], [53] that individuals that die by 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. This may explain why the number of adjectives and adverbs are lowest in GSN notes and highest in DL notes.

3.4.3 Affect Analysis

The analysis of emotions in suicide notes and last statements have often been addressed in research [64], [65]. Table 3 shows sentiments and emotions that were detected for the datasets using LIWC. Overall the highest amount of affect words are in DL notes, whilst the lowest amount is in blog posts. This may also relate back to the level of authenticity usually found in DL notes and lacking in blog posts due to blog posts writers not being as vulnerable.

Type	GSN	NEU1	NEU2	DL1	DL2
Affect	8.92	5.90	5.84	7.78	8.06
Positive emotion	5.69	3.91	3.82	2.97	3.01
Negative emotion	3.16	1.95	1.95	4.72	4.97
Anxiety	0.28	0.27	0.22	0.65	0.67
Anger	0.62	0.68	0.68	1.24	1.35
Sadness	1.06	0.38	0.39	1.74	1.86

Table 3: LIWC Affect Analysis

The highest amount of *Negative emotions* are also highest in DL notes and lowest in blog posts, similarly as before this may refer back to the Tone used in those type of corpora where a higher amount of negative emotions are often correlated with the tone used in a note. Positive emotions are highest in GSN notes, whilst they are lowest in DL notes. This has been found previously by [31], who have found that emotions such as *'love'* are more frequently found in GSN notes than other corpora. Blog posts display the lowest amount of emotions such as anger, sadness or anxiety, whilst this is highest in the DL corpus. It has been shown in prior research that these emotions are more prevalent in DL note writers as these are typical feelings expressed when people suffer from depression [31].

3.4.4 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 GSN and the lowest in DL. Furthermore GSN notes tend to speak most about family relations and least about friends.

Type	GSN	NEU1	NEU2	DL1	DL2
Social processes	11.87	7.93	8.42	8.10	8.10
Family	1.11	0.33	0.40	0.49	0.42
Friends	0.71	0.31	0.44	0.62	0.56

Table 4: LIWC Social Processes

The term *Cognitive processes* encompasses a number of different aspects, where it was found that the highest amount of cognitive processes was in DL notes and the lowest in blog posts (see Table 5).

Type	GSN	NEU1	NEU2	DL1	DL2
Cognitive Processes	12.63	10.41	10.63	16.31	16.30
Insight	2.46	2.09	2.12	3.83	3.54
Cause	1.07	1.47	1.34	2.09	2.06
Tentativeness	2.65	2.52	2.63	3.38	3.66

Table 5: LIWC Psychological Processes

[66] 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. *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 [53], 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. Similar results can be found in our own data, where both Insight and Cause are low in GSN notes, but high in DL notes. *Tentativeness* refers to the language use that indicates a person is uncertain about a topic and uses a number of filler words. It has been argued that participants who use more tentative words, may have not expressed an event to another person and therefore have not processed an event yet and it has not been formed into a story [56]. The amount of tentative words used in DL notes is highest, whilst it is lowest in GSN notes. This might be due to the fact that GSN writers already had to reiterate over certain events multiple times and have made their decision [53].

3.4.5 Personal Concerns

Personal Concerns refers to the topics most commonly brought up in the different notes (see Table 6), where we note that *Work* is most often referred to in NEU notes and lowest in GSN notes, which could be due to blogging often being used for marketing and advertising [67]. *Money* is most often referenced in GSN notes and lowest in DL notes, where this might be due to the fact that [68] lists these two topics as some of the most common reasons for a person to die by suicide. *Religion* is most commonly referenced in GSN notes and lowest in DL notes, where [57] 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. Furthermore the references to *Leisure* are highest in the NEU corpus and lowest in GSN notes. References to *Home* were highest in GSN notes and lowest in NEU notes, which might be due to GSN writers often leaving instructions behind [29], which could references places within a house.

Type	GSN	NEU1	NEU2	DL1	DL2
Work	1.26	1.96	1.96	1.70	1.50
Money	0.67	0.36	0.49	0.40	0.35
Leisure	0.54	1.56	1.51	0.67	0.77
Home	0.46	0.39	0.39	0.48	0.41
Religion	0.68	0.39	0.32	0.14	0.12
Death	0.73	0.14	0.17	0.36	0.57

Table 6: LIWC Personal Concerns

3.4.6 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 [56]. Table 7 shows that the focus of DL notes is primarily in the past whilst GSN and NEU notes focus on the future. The high focus on the past in DL 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 in GSN notes could be due to the writer leaving behind instructions for others [10]. *Relativity* refers to references to space, motion and time in a note.

Type	GSN	NEU1	NEU2	DL1	DL2
Focus past	3.37	3.21	3.46	4.14	3.71
Focus present	14.14	11.15	10.65	14.97	15.67
Focus future	1.89	1.72	1.54	1.22	1.44
Relativity	10.72	13.34	12.95	13.40	13.16

Table 7 : LIWC Time orientation

3.5 Cohen's d effect size

Cohen's d effect size was used to calculate the pairwise importance [69] of each feature. An effect over *d=0.2* (highlighted blue) indicates a small effect, *d=0.5* (highlighted green) indicates a medium effect and *d=0.8* (highlighted yellow) shows a large effect. Furthermore, [69] argued that an effect size of *d=0.5* or higher should be easily seen by humans in real-world examples. It can be seen in Table 8, that most features have a small effect (36.48%), whereas both medium and large effects make up 22.97% and 6.08% of the features respectively and should be clearly visible when examining any posts or notes. Furthermore, it can be seen that categories such as *Dimensional Analysis*, *Affect* or *Function words* show a medium to large effect size across its subcategories, whereas *Cognitive Processes* seem to only have a small to medium effect size for GSN to DL pairwise comparison. Also, it can be seen that features such as *Word per sentence*, *Adjectives* or *Home* do not have any effect on any on the datasets. Other features such as *Clout*, *Tone*, *Anxiety*, *Anger*, *Insight*, *Tentativeness* and *Focus past* do not appear to be important when measuring statistical significance between GSN3 and NEU1 / NEU2 posts. In comparison, there is only one feature (*Leisure*) that is not statistically significant when comparing GSN3 to DL2 / DL3 notes. When comparing GSN3 to DL2 / DL3 notes, the *Affect* category seems to be most important, whereas for a comparison of GSN3 to NEU1 / NEU2 the *Function word* category is most significant. Therefore, it could be argued that in future work, a more fine-grained analysis of sentiment would provide more insight and distinct features to accurately classify suicide notes from depressed notes. On the other hand, for a comparison of suicide notes to 'neutral' posts, a focus on *function* words seems most appropriate. Overall, the category that seems most important across all three datasets is *Function words*, where only one feature (*Negations*) is not statistically significant when comparing GSN3 to DL2.

4 Learning Algorithm

Recurrent neural networks (RNNs) are well suited towards natural language processing tasks due to their ability to handle sequential data [70], 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* [71], 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 [72], [73], [74] in order to overcome this problem.

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 [75]. This allows LSTMs to better learn input sequences and their dependencies and therefore temporal and complex data dependencies are learned on different layers (see Figure 4).

4.1 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:

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

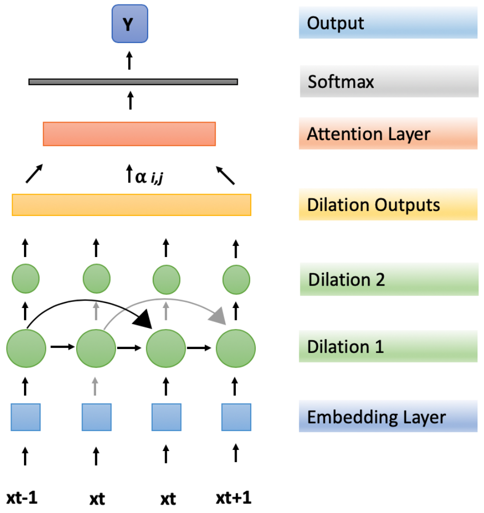
Type	GSN/NEU1	GSN/NEU2	GSN/DL1	GSN/DL2
Word Count	0.207	0.165	0.138	0.24
Word / Sent	0.11	0.03	0.078	0.111
SixItr	0.448	0.61	0.339	0.181
Analytic	0.712	0.861	0.163	0.348
Clout	0.065	0.137	0.808	0.928
Authentic	0.237	0.354	0.64	0.723
Tone	0.036	0.068	0.914	0.962
Function	0.669	0.789	0.217	0.392
Pers. pro.	0.901	1.132	0.248	0.203
I	0.814	1.048	0.21	0.319
Negations	0.658	0.706	0.102	0.3
Verb	0.344	0.627	0.385	0.533
Adjective	0.048	0.04	0.085	0.008
Adverb	0.142	0.236	0.673	0.78
Affect	0.632	0.625	0.239	0.259
Positive emotion	0.438	0.412	0.714	0.754
Negative emotion	0.456	0.508	0.516	0.528
Anxiety	0.014	0.12	0.396	0.367
Anger	0.034	0.015	0.37	0.393
Sadness	0.499	0.396	0.35	0.348
Social proc.	0.605	0.618	0.586	0.678
Family	0.501	0.521	0.38	0.448
Friends	0.333	0.178	0.059	0.122
Cognitive process	0.374	0.402	0.668	0.701
Insight	0.151	0.092	0.554	0.467
Cause	0.219	0.082	0.629	0.631
Tentativeness	0.053	0.015	0.301	0.371
Focus past	0.045	0.006	0.233	0.112
Focus present	0.464	0.702	0.166	0.397
Focus future	0.067	0.184	0.329	0.107
Relativity	0.451	0.391	0.502	0.493
Work	0.302	0.296	0.2	0.177
Money	0.255	0.304	0.21	0.221
Leisure	0.593	0.584	0.105	0.141
Home	0.059	0.108	0.023	0.045
Religion	0.143	0.257	0.288	0.288
Death	0.437	0.519	0.272	0.11

Table 8: Cohen's d effect size

$s^{(l)}$ 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:

$$s^{(l)} = M^{(l-1)}, l = 1, \dots L. \tag{2}$$

We extend the standard dilated LSTM in two ways for our experiments. The standard 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 [72], [74].



Dilated LSTM with stacked units

Furthermore, we extended the earlier implementation with an attention mechanism inspired by [76], 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:

$$u_{it} = \tanh(O + b_w) \tag{3}$$

$$\alpha_{it} = \frac{\exp\left(u_{it}^T u_w\right)}{\sum_t \exp\left(u_{it}^T u_w\right)} \tag{4}$$

$$s_i = \sum_t \alpha_{it} o. \tag{5}$$

5 Experiments

We conduct two different classification experiments, where in both set ups we use a Maximum Entropy classifier to establish a performance baseline.This is due to its suitability to textual data where conditional independence of the features cannot be assumed. Additionally we chose to benchmark our algorithm against the originally proposed Bidirectional LSTM with attention proposed by [76], 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 = 128, 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. For experiment 1 we use the GSN, DL1 and NEU1 dataset, which gives us an overall dataset size of 633 posts. Due to the small size of the dataset we use k-fold cross validation, where k = 10. Whilst for experiment 2 we use GSN, DL2 and NEU2 datasets, where the overall dataset is 5004 and we split the data into *80%* training, *10%* validation and *10%* test data.

6 Results and Evaluation

In the following section we outline the results for both experiment 1 and 2 and provide an evaluation.

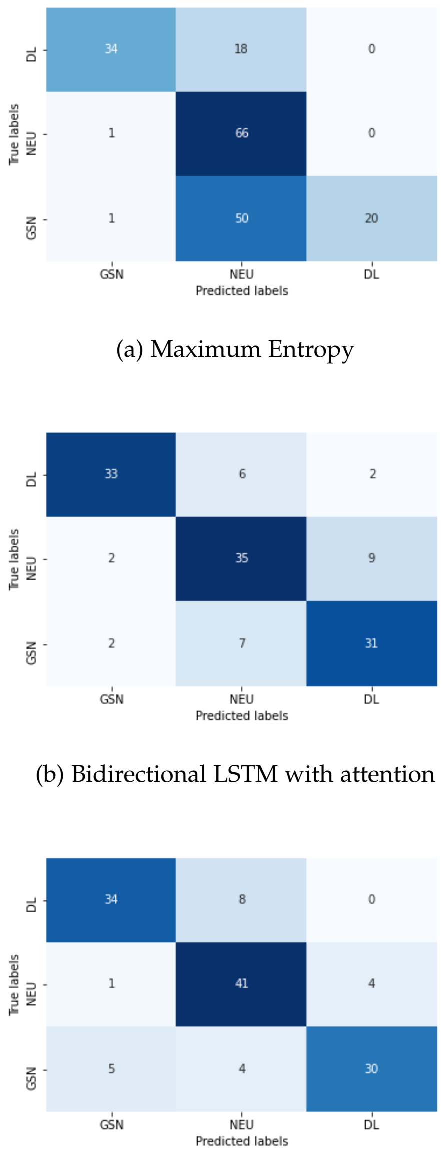
6.1 Experiment 1

All results are shown in Table 9 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 21.93% (Maximum Entropy) and 4% (BiLSTM with attention) respectively. This is due to their ability to handle sequential data of variable length, where as the networks' units decrease hierarchically the information is better retained and different timesteps. 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 given a small dataset, we conducted an additional experiment where the sequence length was restricted to 100. In particular it has been established previously [77], [78] that any vanilla recurrent neural network trained with stochastic gradient descent on a sequence of more than *ten* time-steps will struggle to learn long-term dependencies. 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.

Learning Model	Precision	Recall	F1-score
Maximum Entropy	0.80	0.63	60.73
LSTM (original sequence length)	0.42	0.41	38.05
LSTM (restricted)	0.69	0.66	66.39
BiLSTM	0.75	0.74	74.21
BiLSTM with attention	0.78	0.77	78.10
DLSTMattention	0.82	0.81	81.25
DilatedLSTM ranked units	**0.83**	**0.82**	**82.66**

Table 9: Results of Experiment 2 using precision, recall and f1-score

In Figure 5, three confusion matrices are shown, which demonstrate how well the dilated LSTM with ranked units does compared to the baseline and the best comparable model. Firstly, it has to be noted that in all three figures NEU posts are most accurately classified, then GSN notes and finally DL notes.



(a) Maximum Entropy

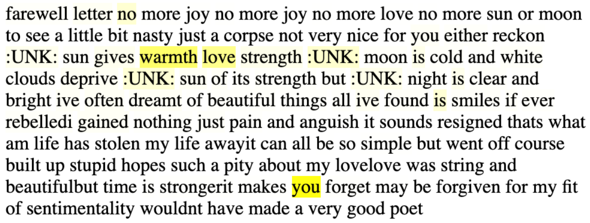
(b) Bidirectional LSTM with attention

(c) Dilated LSTM with ranked units

Figure 5: Three confusion matrices comparing the label classification in three algorithms for experiment 1

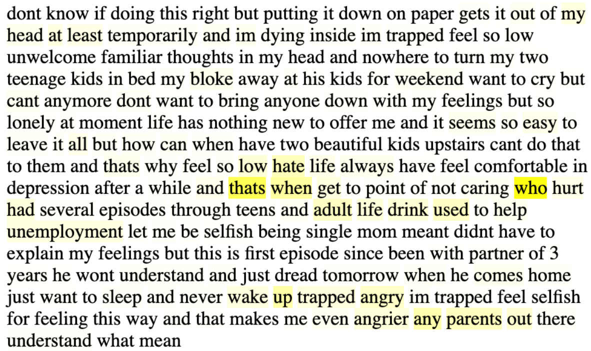
6.1.1 Linguistic Evaluation

In order to see which features are most important to accurate classification we visualise the attention weights of our learning model and show examples from the test set of each dataset (see Figures 6, 7 and 8), where words highlighted in darker shades have higher attention weights. One of the main differences in these three types of documents it is 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. It can also be seen in Table 8 that personal pronouns have a large effect size for GSN/NEU1 and small effect sizes for GSN/DL1. There are a range of different topics and emotions present in each document. More specifically, emotions in GSN notes such as *love*, *joy* and *peacefulness* are present, whilst in DL blogs *anger* and *hate* are predominant. Table 8 also shows that there are small and medium effect sizes for GSN/DL1 comparisons, but fewer effect sizes for GSN/NEU1. This can be seen in NEU1 notes use less emotionally intense language when discussing topics and seem to talk about multiple aspects of a topic. Furthermore the DL1 blog mentions suicidal ideation, however from a linguistic and sentiment perspective it is clearly distinct from a GSN note.



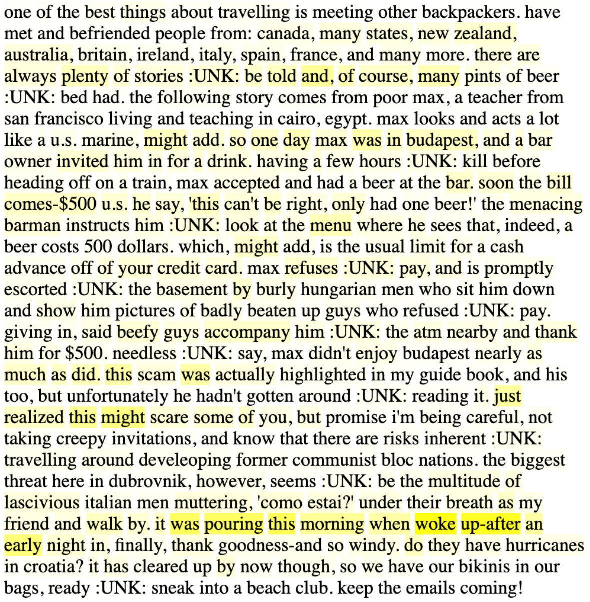
farewell letter no more joy no more joy no more love no more sun or moon
to see a little bit nasty just a corpse not very nice for you either reckon
:UNK: sun gives warmth love strength :UNK: moon is cold and white
clouds deprive :UNK: sun of its strength but :UNK: night is clear and
bright ive often 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 6: Example of a 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 7: Example of a correctly classified 'DL' note



one of the best things about travelling is meeting other backpackers. have
met and befriended people from: canada, many states, new zealand,
australia, britain, ireland, italy, spain, france, and many more. there are
always plenty of stories :UNK: be told and, of course, many pints of beer
:UNK: bed had. the following story comes from poor max, a teacher from
san francisco living and teaching in cairo, egypt. max looks and acts a lot
like a u.s. marine, might add. so one day max was in budapest, and a bar
owner invited him in for a drink. having a few hours :UNK: kill before
heading off on a train, max accepted and had a beer at the bar. soon the bill
comes-$500 u.s. he say, 'this can't be right, only had one beer!' the menacing
barman instructs him :UNK: look at the menu where he sees that, indeed, a
beer costs 500 dollars. which, might add, is the usual limit for a cash
advance off of your credit card. max refuses :UNK: pay, and is promptly
escorted :UNK: the basement by burly hungarian men who sit him down
and show him pictures of badly beaten up guys who refused :UNK: pay.
giving in, said beefy guys accompany him :UNK: the atm nearby and thank
him for $500. needless :UNK: say, max didn't enjoy budapest nearly as
much as did. this scam was actually highlighted in my guide book, and his
too, but unfortunately he hadn't gotten around :UNK: reading it. just
realized this might scare some of you, but promise i'm being careful, not
taking creepy invitations, and know that there are risks inherent :UNK:
travelling around develeoping former communist bloc nations. the biggest
threat here in dubrovnik, however, seems :UNK: be the multitude of
lascivious italian men muttering, 'como estai?' under their breath as my
friend and walk by. it was pouring this morning when woke up-after an
early night in, finally, thank goodness-and so windy. do they have hurricanes
in croatia? it has cleared up by now though, so we have our bikinis in our
bags, ready :UNK: sneak into a beach club. keep the emails coming!

Figure 8: Example of a correctly classified 'neutral' blog

6.2 Experiment 2

The results for experiment 2 can all be seen in Table 10, where we also use precision, recall and f1-score as an evaluation metric. It can be seen in table 10 that the dilated LSTM with ranked units also outperforms the baselines and comparable learning models by more than 10%. Furthermore we note that when establishing a baseline, using the Maximum Entropy classifier, the f1-score is lower than in experiment 1 which reflects how much harder the task is when using an imbalanced dataset. Using the original sequence length on the LSTM in this experiment also shows that there is an improved performance. Overall it can be seen that all neural network approaches outperform the classification results of the baseline and are considerably higher than results from experiment 1. Firstly, this could be due to the increased data size which naturally help neural networks to perform better and secondly it could also be argued that the different learning models find it easier to classify NEU posts due to the imbalance in the dataset.

Learning Model	Precision	Recall	F1-score
Maximum Entropy	0.55	0.67	55.69
LSTM (original sequence length)	0.77	0.71	59.00
LSTM (restricted)	0.78	0.73	64.86
BiLSTM	0.90	0.90	90.43
BiLSTM with attention	0.90	0.90	90.43
DLSTMattention	0.80	0.81	80.70
DilatedLSTM ranked units	**0.96**	**0.96**	**96.1**

Table 10: Results of Experiment 2 using precision, recall and f1-score

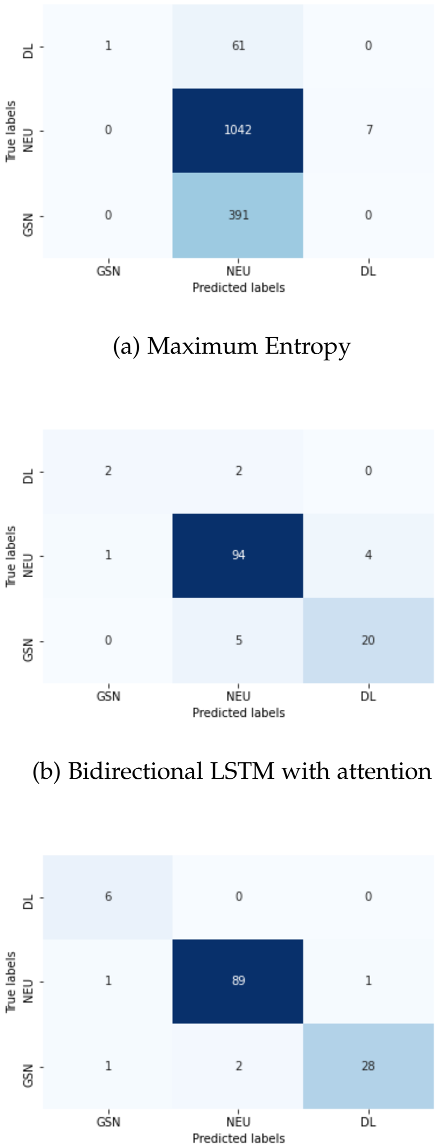
Figure 9 shows three confusion matrices comparing the best performing model to the baseline and the best competing model. Overall it can be seen that the baseline model only classifies NEU posts correctly and only 1 GSN note, whilst it assumes that most DL notes are NEU posts. When comparing the results of the Bidirectional LSTM with attention to the dilated LSTM with ranked units it can be seen that the latter is able to also classify both GSN and DL notes more often. It could be argued that this is due to the learning models ability to access the full sequence length.

6.2.1 Linguistic Evaluation

Figures 10, 11 and 12 all show correctly classified examples of each dataset. It can be seen in the GSN note (see Figure 10), where similar to the findings in the linguistic analysis and for the linguistic evaluation in section 6.1.1. Personal pronouns (*'you'*), positive emotions (*'love'*) and a increased focus on the present (*'is'*) seem to be most important for accurate classification. Similarly in DL2 notes (see Figure 11) references to death (*'im dying inside'*) and work (*'unemployment'*) as well as negative emotions (*'hate'*/*'angry'*) and a increased focus on the past (*'had'*/*'used'*) are assigned the highest attention weights. However, in NEU2 notes (see Figure 12) there seem to be less personal pronouns, increased use of adjectives and adverbs (*'burly'*, *'beefy'* or *'creepy'*) and there seem fewer references to emotions. These findings also correspond with the small to large Cohen's d effect size that was calculated pairwise for each dataset.

7 Conclusion

In this paper we introduced the Dilated LSTM with ranked units and have shown that the learning model is able to successfully distinguish suicide notes from both depressed



(a) Maximum Entropy

(b) Bidirectional LSTM with attention

(c) Dilated LSTM with ranked units

Figure 9: Three confusion matrices comparing the label classification in three algorithms for experiment 2

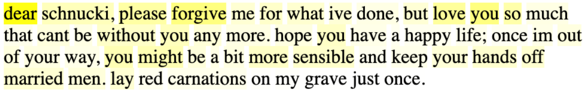

dear schnucki, please forgive me for what ive done, but love you so much that cant be without you any more. hope you have a happy life; once im out of your way, you might be a bit more sensible and keep your hands off married men. lay red carnations on my grave just once.

Figure 10: Example of a correctly classified ’GSN’ note

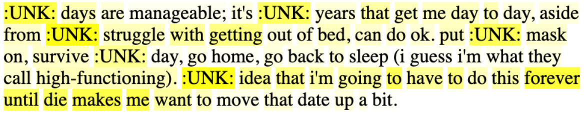

:UNK: days are manageable; it's :UNK: years that get me day to day, aside from :UNK: struggle with getting out of bed, can do ok. put :UNK: mask on, survive :UNK: day, go home, go back to sleep (i guess i'm what they call high-functioning). :UNK: idea that i'm going to have to do this forever until die makes me want to move that date up a bit.

Figure 11: Example of a correctly classified ’DL’ note

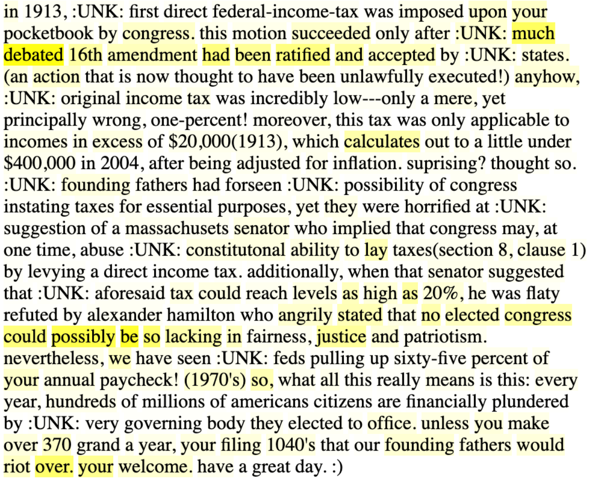

in 1913, :UNK: first direct federal-income-tax was imposed upon your pocketbook by congress. this motion succeeded only after :UNK: much debated 16th amendment had been ratified and accepted by :UNK: states. (an action that is now thought to have been unlawfully executed!) anyhow, :UNK: original income tax was incredibly low---only a mere, yet principally wrong, one-percent! moreover, this tax was only applicable to incomes in excess of $20,000(1913), which calculates out to a little under $400,000 in 2004, after being adjusted for inflation. suprising? thought so. :UNK: founding fathers had forseen :UNK: possibility of congress instating taxes for essential purposes, yet they were horrified at :UNK: suggestion of a massachusets senator who implied that congress may, at one time, abuse :UNK: constitutonal ability to lay taxes(section 8, clause 1) by levying a direct income tax. additionally, when that senator suggested that :UNK: aforesaid tax could reach levels as high as 20%, he was flaty refuted by alexander hamilton who angrily stated that no elected congress could possibly be so lacking in fairness, justice and patriotism. nevertheless, we have seen :UNK: feds pulling up sixty-five percent of your annual paycheck! (1970's) so, what all this really means is this: every year, hundreds of millions of americans citizens are financially plundered by :UNK: very governing body they elected to office. unless you make over 370 grand a year, your filing 1040's that our founding fathers would riot over. your welcome. have a great day. :)

Figure 12: Example of a correctly classified ’neutral’ blog

blogs and ’neutral’ blogs. We have tested the learning model in two different experiment settings, where we have found that accurate classification of suicide notes was easier when the dataset was balanced. However, we have also found that when using the dilated LSTM with ranked units on an imbalanced dataset that makes the overall task more realistic, it was able to identify more suicide notes compared to other learning models. The learning model outperforms the baseline of 60.73% and when using F1-score for evaluation in experiment 1 and achieves and F1-score of 96.1% in experiment 2. Furthermore we have shown that it is possible to achieve better results when significantly reducing the sequences length in a standard LSTM on a small dataset in experiment 1. Therefore demonstrating that accurate classification is possible solely on linguistic patterns in this type of textual data. Therefore these linguistic differences could substantially contribute to future analysis of mental health issues online. 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 and topics discussed in textual data could help further to identify suicidal ideation. Given further research such work could be useful in a number of scenarios, including but not limited to assessing the seriousness of a social media post or suicide attempt in a clinical settings.

8 Acknowledgments

This work was partly supported by the STFC Hartree Centre’s Innovation Return on Research programme, funded by the Department for Business, Energy Industrial Strategy.

References

[1] ITU, “Ai for good summit,” 2019. [Online]. Available: https://aiforgood.itu.int/

[2] WHO, “Sustainable development goal 3,” 2019. [Online]. Available: https://sustainabledevelopment.un.org/sdg3

[3] T. Shioiri, A. Nishimura, K. Akazawa, R. Abe, H. Nushida, Y. Ueno, M. KOJIKA-MARUYAMA, and T. Someya, “Incidence of note-leaving remains constant despite increasing suicide rates,” *Psychiatry and Clinical Neurosciences*, vol. 59, no. 2, pp. 226–228, 2005.

[4] B. Desmet and V. Hoste, "Emotion detection in suicide notes," *Expert Systems with Applications*, vol. 40, no. 16, pp. 6351–6358, 2013.

[5] C. E. Osgood and E. G. Walker, "Motivation and language behavior: A content analysis of suicide notes." *The Journal of Abnormal and Social Psychology*, vol. 59, no. 1, p. 58, 1959.

[6] H. W. Cummings and S. L. Renshaw, "Slca iii: A metatheoretic approach to the study of language," *Human Communication Research*, vol. 5, no. 4, pp. 291–300, 1979.

[7] Facebook, "Suicide prevention," 2019. [Online]. Available: https://www.facebook.com/help/594991777257121/

[8] BBC, "Facebook 'sorry' for distressing suicide posts on instagram," 2019. [Online]. Available: https://www.bbc.co.uk/news/uk-46976753

[9] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, and M. 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*, 2015, pp. 31–39.

[10] J. P. Pestian, P. Matykiewicz, M. Linn-Gust, B. South, O. Uzuner, J. Wiebe, K. B. Cohen, J. Hurdle, and C. Brew, "Sentiment analysis of suicide notes: A shared task," *Biomedical informatics insights*, vol. 5, no. Suppl 1, p. 3, 2012.

[11] R. A. Calvo, D. N. Milne, M. S. Hussain, and H. Christensen, "Natural language processing in mental health applications using non-clinical texts," *Natural Language Engineering*, vol. 23, no. 5, pp. 649–685, 2017.

[12] M. Coulthard, A. Johnson, and D. Wright, *An introduction to forensic linguistics: Language in evidence*. Routledge, 2016.

[13] B. O'dea, M. E. Larsen, P. J. Batterham, A. L. Calear, and H. Christensen, "A linguistic analysis of suicide-related twitter posts," *Crisis*, 2017.

[14] N. Shahreen, M. Subhani, and M. M. Rahman, "Suicidal trend analysis of twitter using machine learning and neural network," in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*. IEEE, 2018, pp. 1–5.

[15] M. Liakata, J.-H. Kim, S. Saha, J. Hastings, and D. Rebholz-Schuhmann, "Three hybrid classifiers for the detection of emotions in suicide notes," *Biomedical informatics insights*, vol. 5, pp. BII–S8967, 2012.

[16] G. Coppersmith, R. Leary, P. Crutchley, and A. Fine, "Natural language processing of social media as screening for suicide risk," *Biomedical informatics insights*, vol. 10, p. 1178222618792860, 2018.

[17] W. Wang, L. Chen, M. Tan, S. Wang, and A. P. Sheth, "Discovering fine-grained sentiment in suicide notes," *Biomedical informatics insights*, vol. 5, pp. BII–S8963, 2012.

[18] J. F. Gunn and D. Lester, "Twitter postings and suicide: An analysis of the postings of a fatal suicide in the 24 hours prior to death," *Suicidologi*, vol. 17, no. 3, 2015.

[19] K. Denecke and Y. Deng, "Sentiment analysis in medical settings: New opportunities and challenges," *Artificial intelligence in medicine*, vol. 64, no. 1, pp. 17–27, 2015.

[20] A. M. Schoene, G. Lacey, A. P. Turner, and N. Dethlefs, "Dilated lstm with attention for classification of suicide notes," in *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, 2019, pp. 136–145.

[21] M. Jaiswal, S. Tabibu, and E. Cambria, ""hang in there": Lexical and visual analysis to identify posts warranting empathetic responses," 2017.

[22] G. Savova, J. Pestian, B. Connolly, T. Miller, Y. Ni, and J. W. Dexheimer, "Natural language processing: applications in pediatric research," in *Pediatric biomedical informatics*. Springer, 2016, pp. 231–250.

[23] C. E. Osgood, "The cross-cultural generality of visual-verbal synesthetic tendencies," *Systems research and behavioral science*, vol. 5, no. 2, pp. 146–169, 1960.

[24] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," *arXiv preprint arXiv:1910.12611*, 2019.

[25] E. S. Shneidman and N. L. Farberow, "Clues to suicide," *Public health reports*, vol. 71, no. 2, p. 109, 1956.

[26] J. J. Shapero, "The language of suicide notes," Ph.D. dissertation, University of Birmingham, 2011.

[27] F. Ren, X. Kang, and C. Quan, "Examining accumulated emotional traits in suicide blogs with an emotion topic model," *IEEE journal of biomedical and health informatics*, vol. 20, no. 5, pp. 1384–1396, 2015.

[28] N. J. Jones and C. Bennell, "The development and validation of statistical prediction rules for discriminating between genuine and simulated suicide notes," *Archives of Suicide Research*, vol. 11, no. 2, pp. 219–233, 2007.

[29] J. Pestian, H. Nasrallah, P. Matykiewicz, A. Bennett, and A. Leenaars, "Suicide note classification using natural language processing: A content analysis," *Biomedical informatics insights*, vol. 3, pp. BII–S4706, 2010.

[30] H. Yang, A. Willis, A. De Roeck, and B. Nuseibeh, "A hybrid model for automatic emotion recognition in suicide notes," *Biomedical informatics insights*, vol. 5, no. Suppl 1, p. 17, 2012.

[31] A. M. Schoene and N. 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*, 2016, pp. 128–133.

[32] L. Chen, A. Aldayel, N. Bogoychev, and T. Gong, "Similar minds post alike: Assessment of suicide risk using a hybrid model," in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, 2019, pp. 152–157.

[33] L. D. Handelman and D. Lester, "The content of suicide notes from attempters and completers," *Crisis*, vol. 28, no. 2, pp. 102–104, 2007.

[34] P. Burnap, G. Colombo, R. Amery, A. Hodorog, and J. Scourfield, "Multi-class machine classification of suicide-related communication on twitter," *Online social networks and media*, vol. 2, pp. 32–44, 2017.

[35] E. Cambria and B. White, "Jumping nlp curves: A review of natural language processing research," *IEEE Computational intelligence magazine*, vol. 9, no. 2, pp. 48–57, 2014.

[36] A. Zirikly, P. Resnik, O. Uzuner, and K. Hollingshead, "Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts," in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, 2019, pp. 24–33.

[37] R. Sawhney, P. Manchanda, P. Mathur, R. Shah, and R. Singh, "Exploring and learning suicidal ideation connotations on social media with deep learning," in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2018, pp. 167–175.

[38] A. Benton, M. Mitchell, and D. Hovy, "Multi-task learning for mental health using social media text," *arXiv preprint arXiv:1712.03538*, 2017.

[39] P. Burnap, W. Colombo, and J. Scourfield, "Machine classification and analysis of suicide-related communication on twitter," in *Proceedings of the 26th ACM conference on hypertext & social media*. ACM, 2015, pp. 75–84.

[40] M. Morales, S. Scherer, and R. Levitan, "A cross-modal review of indicators for depression detection systems," in *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical PsychologyâFrom Linguistic Signal to Clinical Reality*, 2017, pp. 1–12.

[41] P. Resnik, W. Armstrong, L. Claudino, T. Nguyen, V.-A. Nguyen, and J. Boyd-Graber, "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*, 2015, pp. 99–107.

[42] S. Rude, E.-M. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cognition & Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004.

[43] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, and M. Berk, "Affective and content analysis of online depression communities," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 217–226, 2014.

[44] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[45] M. R. Morales and R. Levitan, "Speech vs. text: A comparative analysis of features for depression detection systems," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 136–143.

[46] X. Zhao, S. Lin, and Z. Huang, "Text classification of micro-blog's" tree hole" based on convolutional neural network," in *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*, 2018, pp. 1–5.

[47] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in twitter," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014, pp. 51–60.

[48] D. Mowery, A. Park, M. Conway, and C. Bryan, "Towards automatically classifying depressive symptoms from twitter data for population health," in *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, 2016, pp. 182–191.

[49] H. Binali, C. Wu, and V. Potdar, "Computational approaches for emotion detection in text," in *4th IEEE International Conference on Digital Ecosystems and Technologies*. IEEE, 2010, pp. 172–177.

[50] S. Rosenthal and K. 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*. Association for Computational Linguistics, 2011, pp. 763–772.

[51] A. Bartle and J. Zheng, "Gender classification with deep learning," in *Technical report*. The Stanford NLP Group., 2015.

[52] W.-H. Lin, E. Xing, and A. Hauptmann, "A joint topic and perspective model for ideological discourse," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2008, pp. 17–32.

[53] A. Gregory, "The decision to die: The psychology of the suicide note," *Interviewing and deception*, pp. 127–156, 1999.

[54] I. Pirina and Ç. 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*, 2018, pp. 9–12.

[55] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker, "Effects of age and gender on blogging." in *AAAI spring symposium: Computational approaches to analyzing weblogs*, vol. 6, 2006, pp. 199–205.

[56] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[57] M. A. Just, L. Pan, V. L. Cherkassky, D. L. McMakin, C. Cha, M. K. Nock, and D. Brent, "Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth," *Nature human behaviour*, vol. 1, no. 12, p. 911, 2017.

[58] J. W. Pennebaker, C. K. Chung, J. Frazee, G. M. Lavergne, and D. I. Beaver, "When small words foretell academic success: The case of college admissions essays," *PloS one*, vol. 9, no. 12, p. e115844, 2014.

[59] A. Cohan, B. Desmet, A. Yates, L. Soldaini, S. MacAvaney, and N. Goharian, "Smhd: A large-scale resource for exploring online language usage for multiple mental health conditions," *arXiv preprint arXiv:1806.05258*, 2018.

[60] M. A. Cohn, M. R. Mehl, and J. W. Pennebaker, "Linguistic markers of psychological change surrounding september 11, 2001," *Psychological science*, vol. 15, no. 10, pp. 687–693, 2004.

[61] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying words: Predicting deception from linguistic styles," *Personality and social psychology bulletin*, vol. 29, no. 5, pp. 665–675, 2003.

[62] G. A. Miller, "The science of words," 1991.

[63] J. T. Hancock, C. Landrigan, and C. Silver, "Expressing emotion in text-based communication," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2007, pp. 929–932.

[64] A. M. Schoene and N. Dethlefs, "Unsupervised suicide note classification," 2018.

[65] D. Lester and J. F. Gunn III, "Ethnic differences in the statements made by inmates about to be executed in texas," *Journal of Ethnicity in Criminal Justice*, vol. 11, no. 4, pp. 295–301, 2013.

[66] A. Boals and K. Klein, "Word use in emotional narratives about failed romantic relationships and subsequent mental health," *Journal of Language and Social Psychology*, vol. 24, no. 3, pp. 252–268, 2005.

[67] H. Onishi and P. Manchanda, "Marketing activity, blogging and sales," *International Journal of Research in Marketing*, vol. 29, no. 3, pp. 221–234, 2012.

[68] Mind, "Depression," 2013. [Online]. Available: https://tinyurl.com/y2xhmnf9

[69] J. Cohen, *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[70] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[71] Y. Bengio, P. Simard, P. Frasconi *et al.*, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[72] S. El Hihi and Y. Bengio, "Hierarchical recurrent neural networks for long-term dependencies," in *Advances in neural information processing systems*, 1996, pp. 493–499.

[73] J. Koutnik, K. Greff, F. Gomez, and J. Schmidhuber, "A clockwork rnn," *arXiv preprint arXiv:1402.3511*, 2014.

[74] J. Chung, S. Ahn, and Y. Bengio, "Hierarchical multiscale recurrent neural networks," *arXiv preprint arXiv:1609.01704*, 2016.

[75] S. Chang, Y. Zhang, W. Han, M. Yu, X. Guo, W. Tan, X. Cui, M. Witbrock, M. A. Hasegawa-Johnson, and T. S. Huang, "Dilated recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 77–87.

[76] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. 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*, 2016, pp. 1480–1489.

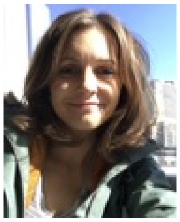
[77] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber *et al.*, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," 2001.

[78] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

Annika Marie Schoene is a PhD candidate in Natural Language Processing at the University of Hull and is affiliated to IBM Research UK. Her research focus is on investigating recurrent neural networks for fine-grained emotion detection in social media data. She also has an interest in mental health issues on social media.

Alexander P. Turner is an Assistant Professor in the Department of Computer Science at the University of Nottingham. His current research interests focus on the application of artificial intelligence techniques in healthcare. He received a PhD in Electronic Engineering from the University of York in 2014. Previously he was a lecturer in the Department of Computer Science at the University of Hull.

Geeth de Mel is a Research Scientist with IBM Research Europe (UK). His research interests are in artificial intelligence—especially in Semantic Web technologies—and on decision support systems in the presence of (or lack of) dynamicity, trust, and provenance. He holds a PhD from the University of Aberdeen, Scotland.



Nina Dethlefs, is a Lecturer in Computer Science at the University of Hull, UK. Her research interests lie at the intersection of natural language processing and machine learning particularly in the areas of natural language generation, interactive systems, text mining and social media, as well as domain transfer and adaptability. She holds a PhD in computational linguistics from the University of Bremen, Germany.