Unsupervised suicide note classification*

Annika M Schoene
University of Hull
A.M.Schoene@2017.hull.ac.uk

Nina Dethlefs
University of Hull
n.dethlefs@hull.ac.uk

ABSTRACT

With the greater availability of linguistic data from public social media platforms and the advancements of natural language processing, a number of opportunities have arisen for researchers to analyse this type of data. Research efforts have mostly focused on detecting the polarity of textual data, evaluating whether there is positive, negative or sometimes neutral content. Especially the use of neural networks has recently yielded significant results in polarity detection experiments. In this paper we present a more fine-grained approach to detecting sentiment in textual data, particularly analysing a corpus of suicide notes, depressive notes and love notes. We achieve a classification accuracy of 71.76% when classifying based on text and sentiment features, and an accuracy of 69.41% when using the words present in the notes alone. We discover that while emotions in all three datasets overlap, each of them has a unique 'emotion profile' which allows us to draw conclusions about the potential mental state that is reflects. Using the emotion sequences only, we achieve an accuracy of 75.29%. The results from unannotated data, while worse than the other models, nevertheless represent an encouraging step towards being able to flag potentially harmful social media posts online and in real time. We provide a high-level corpus analysis of the data sets in order to demonstrate the grammatical and emotional differences.

CCS CONCEPTS

• **Artificial Intelligence → Natural language processing**;

KEYWORDS

sentiment analysis, social media

ACM Reference Format:
Annika M Schoene and Nina Dethlefs. 2018. Unsupervised suicide note classification. In *Proceedings of WISDOM Workshop (WISDOM'18).* ACM, New York, NY, USA, Article 4, 9 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Microblogging has become increasingly popular across a range of different social media platforms. Whilst there has been a lot of focus on opinion mining for consumer products [59] there has also been a surge of sentiment analysis (SA) and emotion detection for mental health purposes [44]. The amount of data created through social media platforms such as Twitter is almost exceeding 140 million tweets a week[1], which allows a unique insight into how Twitter users feel or behave [26]. Other social media platforms such as Facebook have launched online campaigns that encourage its users to report other users they might be concerned about.[2] The main area of focus for many of these campaigns have been depression or suicidal feelings expressed by social media users through posts or tweets.[3] It has been argued by many active in the research area that there is a great need of being able to detect mental states such as suicidal tendencies online [30]. Young people particularly are at risk, with suicide being the second leading cause of deaths for those aged 15 to 29.[4] In addition to that there has been a trend recognised that people leave their suicide notes on social media platforms [11]. Furthermore it has been found that especially people aged 16-24 year old and 25-34 year old are most active on social media platforms like Twitter[23].

It has been argued in previous research that our drive or motivation affects the way in which we communicate and therefore it is believed that our spoken and written language represents those shifting psychological states [51]. This argument has been taken further by [28] who suggested that there is a shift in ones linguistic expression due to the aroused cognitive state suicidal individuals experience. Previous work in analysing suicide notes has focused significantly on identifying hand-crafted features that distinguish suicide notes from either forged suicide notes [43] or other types of text [52].

In this paper we extend the study in [61] and show that similar results can be achieved using deep learning models that work on the word-level alone, i.e. without the requirement for hand-labelled data. We achieve a classification accuracy of 69.41% distinguishing suicide notes, depressive and love notes based only on the words occurring in the notes without further annotation. This is in comparison to earlier work that achieved 86.6 using Logistic Regression and hand-annotated linguistic and sentiment features. We find that using attention and pre-trained word embeddings make a significant contribution to classification accuracy. Furthermore, we show that annotating the data for sentiment features increases accuracy further to 71.76%. Additionally, we find that classifying emotion based on the sequence they occur in each note yields an accuracy of 75.29%, which gives us insight into the emotional journey of the note writer. Finally, we use the genuine suicide notes corpus for comparison in our feature analysis in order to demonstrate the significant grammatical and emotional differences compared to other types of discourse online.

*Produces the permission block, and copyright information

WISDOM'18, August 2018, London, UK

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

[1] TwitterStats
[2] SafeFacebook
[3] FacebookSuicide
[4] WHOStats

2 RELATED WORK

Previous work has explored a range of approaches in classifying sentiment from corpora.

2.1 Sentiment Analysis

In recent years a lot of research has focused on detecting the polarity of textual data within the field of sentiment analysis, mainly aiming to do binary classification by detecting positive or negative sentiment in text [22]. Other work has taken sentiment analysis into other directions such as looking at emotion intensity where the main goal is to classify the intentsity of an emotion felt by user in a tweet [48]. Some of the most commonly used methods for polarity detection include frameworks such as SenticNet [21], SentiWordNet [13] or LIWIC [54]. One of the most popular resource for collecting data from social media is Twitter and a shared task has been dedicated to it with 48 teams participating in 2017 [58]. This trend has also spilled over into other research fields such as psychology. Using Twitter as a resource to detect different mental health conditions has become more relevant over recent years. Some work has focused on detecting mental health signals related to conditions such as bipolar disorder, major depressive disorder, post-traumatic-stress disorder and seasonal affective disorder [25]. In their work [50] have looked extensively at the which features are relevant when classifying depression in tweets.

Statistical approaches to SA encompasses both traditional machine learning techniques such as Support Vector Machines [66] as well as deep learning models [57]. Statistical approaches have been on the rise over recent years due to the popularity and success of deep learning models that have achieved competitive results in a number of SA tasks [58]. Classification accuracy for binary SA tasks such as movie reviews are particularly successful with the highest accuracy achieving 82.2%, when applying traditional machine learning techniques [53]. The reviewers of the annual SemEval shared task [58] have also noticed an increase in deep learning techniques used for SA task, where almost half of the submissions used some form of deep learning. Furthermore the winning team of the overall classification task achieved an accuracy of 68.1% [16] by utilising deep learning techniques [58]. It is important to note that deep learning approaches rely heavily on word-embeddings, which represent words based on their co-occurrence with each other in a vector [65]. Deep learning approaches also rely heavily on annotated and high volume datasets [33]. Therefore it has been argued that these methods do not perform as well on smaller *'linguistic units'* such as sentences or clauses as semantic value cannot always be derived from the frequency of lexical items or their co-occurrence with each other [20].

SA has also been applied in medical settings [29], where research is not only limited to suicide note analysis but also other healthcare settings such as patient's opinions or emotions towards a service or treatment. Overall, research in the field of SA has become increasingly interested in looking at content created online that may solicit need for help [42] or detecting mental health issues [60]. Work by [19] has looked at identifying suicidal ideation on twitter by using lexical, structural and sentiment features. In their study they used traditional machine learning algorithm and achieved an F-measure of 0.728.

2.2 Work on suicide note classification

The analysis of suicide notes has been used in various academic settings such as psychology or forensic linguistics in order to either identify the genuineness of a suicide note or to predict the state of mind of a note writer. It has been argued in previous research that our drive or motivation affects the way in which we communicate and therefore it is believed that our spoken and written language represents those shifting psychological states [51]. This argument has been taken further by [28] who suggested that there is a shift in one's linguistic expression due to the aroused cognitive state suicidal individuals' experience.

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 [27]. Another field where the analysis of suicide notes is crucial is psychology, where one of the most commonly cited studies has been conducted by [64]. 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 [63] and researchers such as [52] have compared this set of suicide notes with a set of normal letters to friends. These studies have been taken further recently by [55] who also used this set of suicide notes and hypothesised that when applying the set to a machine learning algorithm it would outperform mental health professionals in classifying suicide notes correctly. Amongst other things they annotated the suicide notes with emotions and found that the machine learning algorithms were classifying the suicide notes as well as humans. It is predicted by [55] that the results of this study and further studies can be used in various decision making settings such as clinical assessment of suicide attempters when being admitted to hospital as well as to eliminate 'malingerers who feign psychiatric illness for ulterior motives'. Furthermore it has been argued by [30] that there is a need for 'automatic procedures that can spot suicidal messages and allow stakeholders to quickly react to online suicidal behaviour or incitement'. Work by [36] has conducted a qualitative analysis on 145 tweets that were posted 24 hours prior to a girls death on the popular micro-blogging platform Twitter. In their study they used a number of different linguistic and sentiment measurements and argued that this kind of work 'permits to study the mind of suicides as they approach their act. It allows us to monitor short-term indications of future suicidal behaviour'.

Over the years there has been much research conducted into the accurate classification of suicide notes, with most of them using traditional supervised machine learning methods. Particullary the work of [56] has been influential in the field and in their study theyt have found that there are fifteen different emotional concepts which prove to be significant in identifying genuine suicide notes. These fifteen emotions will be used as a guideline for the sentiment classification experiment in this article. These fifteen sentiment features have also been used by [67] in the i2b2/VA/ Cincinnati Medical Natural Language Processing Challenge. The aim of the

challenge was to develop a model which could automatically identify emotions on sentence level of a suicide note. The hybrid model developed by [67]) achieved an accuracy of 61.39% in detecting emotions using various techniques such as machine learning based emotion classification. The result of this experiment shows that it is possible for a machine to correctly identify emotions in suicide notes. It is argued in their paper that one of the key factors for successful identification of emotions is to split the 15 pre-specified emotions into three different classes. More recent work has focused on combining both sentiment and linguistic features which led to the applied machine learning accuracy to achieve an accuracy of 86.6% [61].

The following experiment series aims to explore whether it is possible to use unannotated and unsupervised learning models and still accurately distinguish three types of corpora from each other.

3 DATA COLLECTION

Three different corpora were used for our experiments, which were taken from three different sources. 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.

Genuine Suicide Notes (GSN) This corpus had been collected from various sources including newspaper articles and already existing corpora from other academic resources such as [64], [45] and [31]. The copies obtained from online sources such as newspapers were only used if a full copy of the genuine note was present. The suicide notes collected from online sources included [9] , [10], [7], [1], [8] and [3].

> *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.

Love Posts (LH) This datasets has been collected from public post of the Experience Project website. The love corpus was collected from the thread 'I Think Being In Love Is One Of The Best Feelings Ever' [6] and 'I Smile When I Think Of You [5]. This corpus was chosen as it could be argued that this corpus would demonstrate the opposite type of emotions used in the GSN corpus, which could also lead to a change in the language used within the posts.

> *My boyfriend and I are in love with each other and it's such a wonderful feeling. He told me that he's in love with me and it was such a wonderful overwhelming feeling that it made me cry tears of joy. I love him so much.*

Figure 2: Example of a love note.

Depression (DL) This dataset was also collected from the public section of the Experience Project website. The depression corpus was collected from the group 'I Fight Depression And Loneliness Everyday [4]. The Depression corpus was collected as it may be close in the emotions and language usage to the suicide note corpus.

Corpora	GSN	LH	DS
Word count	137.77	65.12	112.37
Av. Note Length	144.60	70.78	120.85
Av. Sentence Length	15.0	12.0	15.0
Allness Terms	123	85	113
Cognitive Processes	12.36	15.05	16.95
Characters	773.97	340.35	614.63
Nouns	27.29	10.00	16.25
Verbs	25.83	12.96	23.28
Adjectives	7.23	4.11	6.56
Adverbs	8.01	4.88	9.86
Pronouns	20.13	10.57	15.80
Lexical Diversity	6.56	6.15	7.13

Table 1: Linguistic Features across all three corpora

> *I've stopped taking my meds and its all hitting me hard. I thought by taking those meds it would make me better but it didn't, it just made me worst! ! I am done with everything*

Figure 3: Example of a depressed note.

4 CORPUS ANALYSIS

The following sections aims to give a better insight into the three corpora and show how previous work has used hand-crafted features for distinguishing suicide notes from other corpora.

4.1 Linguistic Features

Suicide note research has not only focused on the sentiment conveyed in notes, but also on linguistic[52] and content [37] features. Research conducted by [43] 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 [61] 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). This result was also benchmarked against a J48 decision tree (78%) and a Naive Bayes classifier achieving an accuracy of 74%. The following analysis gives an insight into the linguistic composition of our three different corpora and how the content relates to previous findings.

Average note length. In Table 2 we can see the average number of grammatical and content features for each post or note for each dataset. The first feature to be analysed will be the length of the each individual note collected as it is argued by [35] that suicide notes are greater in length due to the fact that the suicidal individual wants to convey as much information as possible. This is due to the note writer's feeling that they will not have time to convey this information at a later point [35]. Table 2 also shows the overall length of each corpus which is then divided by the number of

notes collected in order to compute the average length of each note. Research by [35] proposes that because there is a high amount of important information a suicidal individual wants to write down before their imminent death, which results in an overall greater length of communication. Another feature demonstrated in Table 2 is the length of communication overall in all three corpora. This observation proves to be true for the three corpora analysed as there is a significant difference between the lengths of the three corpora. In addition to that it can be seen that the corpora differ significantly in the average length per note and the notes of the GSN corpus is almost double in length compared to the LH corpus. When looking at the word count for the four different corpora it becomes apparent that there the GSN corpus has by far the highest average word count and length per note. Research by [35] gives two main reasons for the greater length of communication: Firstly he argues that suicide notes are greater in length due to the fact that the suicidal individual wants to convey as much information as possible. Secondly it is believed that the note writers feels that they will not have time to convey this information at a later point.

Average Sentence length. The next feature to be analysed is the average sentence length (ASL) of a note and the software LIWC [54] has been used for analysis. It is argued by [52] that in a genuine suicide note the ASL is shorter and that there is a higher focus on conveying the most important facts. Therefore it has been observed that there is usually a small amount of adjective or adverbs in genuine notes [52] . A similar observation was made by [49] who argues that in a higher cognitive state such as high levels of arousal a person focuses on providing only essential information. Table 2 compares the ASL across all three corpora. The GSN corpus and the DL corpus have scored the same results in testing for ASL. One explanation for this may be found when looking at [12] who explains that it has been proven for a long time in clinical settings that there is a similarity between the state of mind of a suicidal person, who can also experience depression. When comparing the LH corpus to the other two corpora it becomes clear that although the number of tokens in the corpus is smaller, the sentence length is almost as high as the one of the GSN and DL corpora. It could be argued that this phenomenon may be due to a higher amount of adjectives used in a sentence, which will be tested at a later point. In addition to this, it has been argued that people who communicate under stress tend to break their communication down into shorter units [52]. The research however suggested that there is no significant difference in the overall length per unit when comparing suicide notes to regular letters to friends and simulated suicide notes [52].

Nouns. We use NLTK's Part-of-Speech Tagging to identify the linguistic characteristics [18]. Research has suggested that individuals who commit suicide tend to use more nouns and verbs in their notes [35] and only a small amount of adjectives and adverbs [52]. In addition to that [43] state that a person who is going to commit suicide is under a higher drive and therefore it is more likely for them to reference a high amount of objects (nouns) compared to any other type of word such as verbs. This has also been supported by other studies which found that under higher degrees of stress the ability to retrieve nouns tends to stay the same whereas the retrieval of verbs was less successful [38]. When looking at the other three corpora for comparison it can be seen that this is not true for either the LH or DS corpus, however it is true for the Twitter corpus. Therefore it could be argued that the people whose posts have been collected for the DS and LH corpus are under a lesser degree of stress. Another reason why this might be the case is that the amount of verbs is only higher than the amount of nouns in the LH and DS corpus.

Verbs. Another feature to be included is the number of verbs used in the three corpora. Table 2 compares the average number of verbs per note in each corpus. Although [43] argue that there is a high number of verbs in the genuine suicide notes compared to forged notes and it can be seen that the GSN corpus has the lowest amount of verbs present. As the number of verbs is compared to the number of adjectives and adverbs in [43] study, it can be seen that their findings concur with the ones for this corpus as in every corpus the number of verbs is higher than the adjectives and adverbs on their own or combined; however it has to be noted that this proves to be true for all three corpora.

Adverbs and Adjectives. Adverbs have been defined as [39] as a 'lexical modifier of a non-nominal head'. It could therefore be argued that seen as the number of verbs is already decreasing in people who would like to commit suicide tend to not use many adverbs either. On the other side adjectives are used to modify nouns [15] The highest in amount of adjectives can be found in the GSN corpus whereas the lowest is found in the Twitter corpus. This might be due to the fact that there is already limited space to convey a message in a single Tweet and therefore people tend to use less amplifying language.

Allness Terms. Furthermore it has been suggested by [51] that when a person is a highly emotional they tend to polarise or communicate points in a more extreme manner. They have called these words 'allness terms' and therefore the usage of the following terms (Table 2) has been explored in the three corpora using [18]. It can be seen that the GSN corpus has the highest amount of allness terms, whereas the LH corpus has the lowest. These findings concur with the previously mentioned study by [52] and it could be argued that the DL corpus is scoring in the middle because people who wrote in it may be in an emotionally similar state, but still not to the same extent as writers of suicide notes.

Cognitive Processes. Another feature has been proposed by [35], who suggested that there is a lower amount of cognitive processes identifiable in a genuine suicide note as the writer has already finished the decision making process. Moreover [35] argued that this was done by identifying the number of cognitive process words, which would be higher in simulated notes as the writer would still try to justify his or her choice. Therefore [54] has been used in order to identify the cognitive process in every individual note of each corpus. For comparison purposes the results of each corpus have been added up and then divided by the number of notes collected to identify the average amount of cognitive activity per note (Table 2). It can be seen that the GSN corpus has the lowest cognitive activity per note compared to the other two corpora. The highest amount of cognitive processes was found in the DL corpus and it could be argued that this is due to the note writer is still in a process of making a decision or evaluating a situation. This analysis validates [35] 's theory and also suggests that there is a significant

difference between notes of the GSN and DL corpus in terms of cognitive activity. Furthermore [41] found in their study that there is a lower amount of cognitive processes in suicide notes compared to simulated suicide notes.

Pronouns. Furthermore it has been stated by [46] that there are more references to other people in genuine suicide notes. This was measured by using pronouns in order to determine whether there are any significant differences. All personal pronouns will be used for comparison with NLTK. The personal pronouns referring to oneself are marked as 'self', pronouns referring to other people will be called 'other' and those referring to a group in the first person plural will be classed 'both'. As it can be seen in Table 2 there is a higher reference in each corpus to oneself compared to others. Therefore the findings of [46] do not prove to be true for the data collected in this dissertation. It could be however argued that due to the amount of pronouns used in the three corpora that there are differences which may be useful to further investigate in order to analysis whether the overall usage of pronouns is significant on an individual note level. The average amount of pronouns used per note in each corpus proves to be significantly different as the usage in the DL corpus is almost half of the amount used in the GSN corpus (Table2).

Lexical Diversity. Another feature described by [52] in their study showed how they used Type/Token Ratio (TTR) to discriminate between suicide notes and regular letters written to friends. TTR is computed by dividing the number of individual words by the number of total words in a corpus. They have found that due to the heightened cognitive state of mind of a suicidal person, there should be a lower TTR ratio in suicide notes than in letters to friends. This process has been called Lexical Diversity in [18] and has been used to compute it for all three corpora. Although it has been argued previously that people intending to commit suicide can demonstrate similar behavioural patterns like people who suffer from depression, it can now be seen that there is a significant difference in how many different words they use in a note (Table 2). It can also be seen that the LH corpus is less lexical diverse than the GSN and DL corpora. One explanation for this phenomenon could be due to the fact that the LH corpus is shorter in length and therefore the number of individual words is lower compared to the other two corpora.

Corpora	GSN	LH	DS
Work	1.28	0.49	0.97
Leisure	0.55	0.31	1.03
Home	0.53	0.22	0.51
Money	1.45	0.27	0.30
Religion	0.88	0.3	0.09
Death	0.74	0.01	0.64

Table 2: Average reference to topic per note

4.2 Sentiment Features

Substantial research has been conducted on the emotional words of people who contemplate suicide or have committed suicide [46]

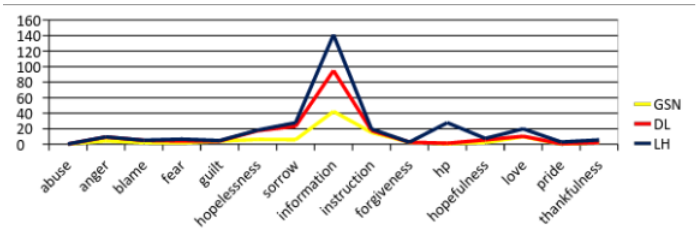

Figure 4: Distribution of Emotion across all corpora.

over many years. This research also included the sentiment conveyed in suicide notes and [56] has argued that a total of fifteen different sentiment features are most significant to suicide notes. Table 1 shows the number of sentiment features occurring in H% of in each of the dataset, whilst Figure 11 shows a distribution of all features across all corpora.

Corpora	GSN	LH	DL
Instruction	15.23	1.93	2.39
Information	41.55	44.74	53.85
Anger	4.65	0.35	5.05
Fear	1.55	2.11	2.94
Blame	2.00	0.18	2.94
Hopelessness	6.26	0.70	10.37
Abuse	0.13	0.18	0.09
Sorrow	5.94	4.56	17.06
Guilt	4.39	1.93	0.09
Thankfulness	2.26	2.98	0.28
Forgiveness	2.65	0.00	0.09
Hopefulness	2.32	1.93	3.03
Love	10.13	9.30	0.55
Pride	0.32	2.11	0.46
Happiness	0.65	27.02	0.83

Table 3: Sentiment features in % across all three corpora

There are some general observations to be made about the emotion distribution in the three different corpora. It can be seen in Figure 4 that all emotions are present in the GSN and DL corpus, however not all emotions are present in the LH. It could be argued that the incompleteness of all features in the LH corpus is to be expected as the sentiment features were designed for a specific domain. However, it is interesting to note that all sentiment features are present in the GSN and DL corpus, which could support the hypothesis that sentiments in both corpora are close. An important observation has been made by [45], who argues that there is a greater confusion of emotions in suicide notes. This may be one of the reasons why the different emotions, occur with a higher percentage in the GSN corpus and less or not at all in the other two. [2] describes on their website typical feelings people experience when suffering from depression such as hopelessness, sorrow as well as anxiety. These emotional concepts match the ones primarily found in the DL corpus and therefore it could be argued that overall the emotions found in the individual corpora demonstrate that the collected notes reflect the purpose of each individual corpus.

Information is the one feature which is the largest in percentage and present in all three corpora. This may be due to the fact that the clauses labelled with the concept of 'information' are mainly descriptive and let the reader know things such as where a specific item is placed [67]. Furthermore the results of the GSN corpus correspond to the findings of [46], who argue that there is a high likelihood that a person leaves instructions behind for the survivors. Additionally [32] has found that 60% of people convey their love for those who they leave behind in a suicide note, which would explain why the emotional concept of love is so prominent. On the other hand it has been argued by [17] that the emotions happiness and love are closely related to each other because sharing intrinsically valuable activities with a loved one can generate happiness on both sides. Therefore it could be argued that it is logical that besides the feature information, love and happiness are the two best performing emotions in the LH corpus. Additionally, it has been found by [36] that positive emotions increase when a person is close to committing suicide. The concept of forgiveness is not present in the LH corpus, which might be due to the fact that people in this domain mainly write about positive experiences and therefore are less likely to be in a position where they would need to apply it. Figure 4 demonstrates the similarities and differences of the corpora when compared.

5 LEARNING MODEL

Our primary model in this study is the Long short-term memory (LSTM) given its suitability for language and time-series data [40]. We compare a standard LSTM with an LSTM that adds a bidirectional layer and attention. An illustration of our model is shown in Figure 5. We feed into the LSTM an input sequence $\mathbf{x} = (x_1, \ldots, x_N)$ of words in a note alongside a label $y \in Y$ denoting a dataset from ['GSN', 'DE', 'LOVE']. The LSTM learns to map inputs x to outputs y via a hidden representation $\mathbf{h}_t$ which can be found recursively from an activation function

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

where t denotes a time-step. Examples of activation functions are sigmoid, tanh or ReLU, the best tends to differ per learning task. During training, we minimise a loss function, in our case categorical cross-entropy as:

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

LSTMs manage their weight updates through a number of gates that determine the amount of information that should be retained and forgotten at each time step. In particular, we distinguish an 'input gate' i that decides how much new information to add at each time-step, a 'forget gate' f that decides what information not to retain and an 'output gate' o determining the output. More formally, and following the definition by [34], this leads us to update our hidden state $\mathbf{h}$ as follows (where σ refers to the logistic sigmoid function and c is the 'input gate'):

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

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

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

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

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

Bidirectional/Attention. Despite their success at modelling sequential data, LSTMs typically predict outputs only from past sequences. To address this limitation, we add a bidirectional layer to our model which concatenates the forward state and the backward state into a single vector [62], thus considering the best output prediction based on both $x_0 \ldots x_{t-1}$ as well as $x_{t-1} \ldots x_0$. We follow the implementation by [68], who argues that not all words are equal to a sentence meaning and therefore introduces an attention mechanism. The use of attention has yielded significant increase in document classification accuracy for a number of tested datasets [68]. We used [14] attention mechanism, which allows access to a weighted distribution over the input state during prediction making in addition to the hidden state and often leads to better results for longer sequences.

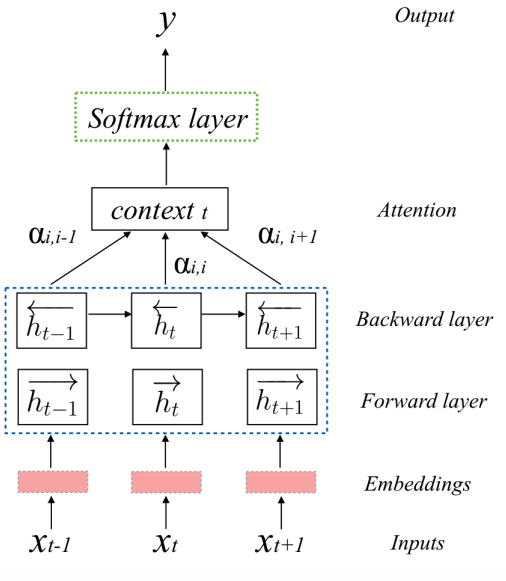

Figure 5: BiLSTM with Attention.

6 EXPERIMENTS AND RESULTS

For our experiments we used word-embeddings [47] as our input features into the learning model. All word-embeddings were pre-trained on our own dataset with the word embedding dimension set to 100 and the maximum number of most common words was set to 5000. Furthermore, we use 80% of our data for training and 20% for validation. All sentiment features are the same features discussed in section 4.1. All experiments were conducted using Keras [24], including a custom attention layer. To assess the importance of different features, we design three experiments:

(1) Classification based on word features only.
(2) Classification based on sentiment and word features.
(3) Classification based on sentiment features only.

Figure 6 shows mock examples of the data used in the experiments. All result are summarised in Table 4. As we can see, the best

(1) "I hate dogs, but love cats."

(2) " I hate dogs [hate], but love cats [love] ."

(3) " hate, love "

Figure 6: Mock examples of the data used in the three Experiments

results are achieved by the biLSTM model with attention based on sentiment features only. Classification from text and sentiments is second best, while text only is only slightly worse than using both text and emotions. Overall it therefore seems that hand-labelled emotions are important to make accurate predictions, however, classification from unlabelled data is still significantly better than a majority baseline of 33.33%. In the following, we provide an analysis of the result on text and emotions and emotions only to shed some light on relevant features and patterns.

Experiments text and emotion features. The results below show that learning with unannotated data is possible (see Table 4), achieving an accuracy of 55.00% with a vanilla LSTM and 75.95% with the bi-directional LSTM with Attention. These results are especially encouraging as to the best of our knowledge no previous research has been conducted that has avoided any kind of data annotation for suicide note classification. The results achieved with the Bi-directional Attention LSTM are encouraging as they are exceeding those of [67], who achieved 61.39% accuracy using traditional machine learning techniques in a hybrid model, such as Support Vector Machines and Naive Bayes. This result also shows that emotion features are important and relevant to more accurate classification of suicide notes. Figures 7, 8, and 9 show example predictions for each category illustrating prediction making.

i'm so tired of feeling sad [sorrow]

Figure 7: Prediction example for DL corpus.

it's like a shot of uber confidence i always feel better about myself for making someone happy just by being me [happiness]

Figure 8: Prediction example for LH corpus.

Experiments with emotion sequences. This experiment was inspired by the observations made in [36], where qualitative reports indicated that positive emotions increase the closer a person gets to committing suicide. Therefore we removed all textual data from the data and only used emotion features in the sequence that they occur within a note. This hypothesis might also explain why emotions such as love are occurring more frequently in the GSN corpus compared to the other two corpora. In order to further investigate this

dear Elinor [information] the reason for my despondency is that you'd prefer the company of almost anyone to mine [anger] you told me you had nothing to look forward to on week ends [blame] you told me you preferred living alone [blame] ...

Figure 9: Prediction example for GSN corpus.

phenomenon we chose ten random notes from each corpus and visualized the emotion sequences using heatmaps. Figure 9 describes the key we used to represent emotion categories numerically, where all light coloured emotions could fall into the category 'negative' and all darker coloured emotions fall into the category 'positive'. Due to the varying length of individual notes a place-holder has been assigned to the number 16.

Figure 10: Key for emotion labels in heatmap.

The results displayed in Figure 10 show for the DL corpus that there is not a large variety of different emotions in the selected examples and would mostly fit into the category of 'negative' emotions such as sorrow or anger. However, due to the nature of the notes collected for this corpus it was to be anticipated that most emotions would be negative. A similar principle applies to the results shown for the LH Corpus, where the majority of emotions could be labelled 'positive', with only a small amount of variation throughout the notes. Finally, the results for the GSN demonstrate that there could be some evidence for the hypothesis made by [36]. Many of the emotions that appear towards the end of a note could be labelled 'positive', whereas the lighter colours at the start of each note indicate more 'negative' emotions.

7 CONCLUSION

Overall our results show that it is possible to accurately classify both unannotated data and emotion category annotated data using unsupervised deep learning methodologies and achieve accuracies that are close to results achieved by traditional machine learning algorithm. This may lead to some interesting discussion about how important hand-crafted features are overall when considering the amount of time and money is used to develop these. However, it is important to still explore these features as our analysis of sentiment and linguistic features have also shown that there are significant

Model	Text only	Text and emotion	Emotion Sequences
Majority	33.33%	33.33%	20.46%
Vanilla LSTM	53.77%	50.00%	57.55%
BiLSTM with attention	69.41%	71.76%	75.29%

Table 4: Experiment results comparing a vanilla LSTM with a bidirectional LSTM with attention for classification from text only, emotions only and text and emotions. All results are in accuracy in %.

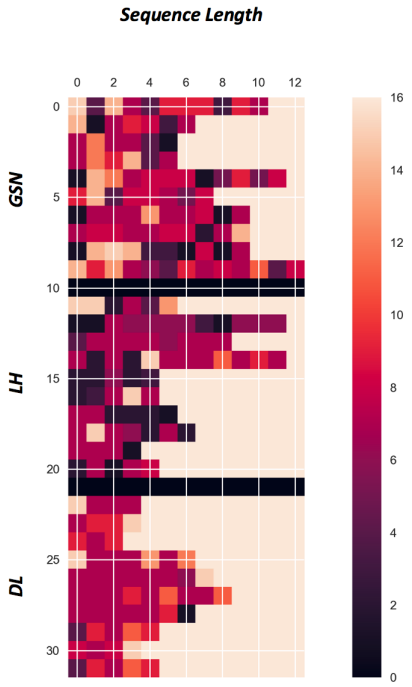

Figure 11: Heatmap illustrating emotions in examples from DL, LH and GSN corpora.

differences when comparing suicide notes to other types of discourse. These features might give us more conclusive proof of what makes a suicide note and how further exploration of especially sentiment features could lead to more accurate identification of suicidal ideation. In addition to this our experiments have shown that there sentiment features contribute to the accurate classification of suicide notes. Furthermore, we have shown that one can accurately classify just the emotion labels used in the data. We used heat-maps to further shed light on this point and provided initial evidence to the hypothesis. This could mean that there is some importance in how often and in which order emotions occur in suicide notes. In addition to this we have also observed that classification accuracy can be improved by using more tailored learning models.

8 ACKNOWLEDGMENTS

We acknowledge the VIPER high-performance computing facility of the University of Hull and its support team.

REFERENCES

[1] [n. d.]. A letter from Angel Green's room. http://archive.jconline.com/article/20130318/NEWS12/303180021/A-letter-found-Angel-Green-s-room. Accessed: 2015- 03- 17.
[2] [n. d.]. Depression. https://www.mentalhealth.org.uk/a-to-z/d/depression.
[3] [n. d.]. I don't want to die but I want to be saved from the pain': Mother shares heart-wrenching suicide note of bullied daughter, 14, as she seeks to get anti-bullying bill passed. . http://www.dailymail.co.uk/news/article-2303946/Angelina-Greens-mother-shares-heart-wrenching-suicide-note-seeks-anti-bullying-passed.html.
[4] [n. d.]. I Fight Depression And Loneliness Everyday. http://www.experienceproject.com/groups/Fight-Depression-And-Loneliness-Everyday/25529.
[5] [n. d.]. I Smile When I Think Of You. http://www.experienceproject.com/groups/Smile-When-I-Think-Of-You/158650.
[6] [n. d.]. I Think Being In Love Is One Of The Best Feelings Ever. . http://www.experienceproject.com/groups/Think-Being-In-Love-Is-One-Of-The-Best-Feelings-Ever/651982.
[7] [n. d.]. Indiana girlâĂŹs public suicide and heartbreaking note sparks anti-bullying legislation in the state. http://www.nydailynews.com/life-style/health/indiana-girl-suicide-heartbreaking-note-spark-anti-bullying-legislation-article-1.1308060.
[8] [n. d.]. Lita Broadhurst. http://solarpix.photoshelter.com/image/I0000NGqLyhxzqPg.
[9] [n. d.]. MS Police release suicide note left by British teenager who killed US policeman. http://www.telegraph.co.uk/news/worldnews/northamerica/usa/10722357/Police-release-suicide-note-left-by-British-teenager-who-killed-US-policeman.html.
[10] [n. d.]. My birth is my fatal accident: Full text of Dalit Student RohithâĂŹs suicide letter. http://indianexpress.com/article/india/india-news-india/dalit-student-suicide-full-text-of-suicide-letter-hyderabad/.
[11] [n. d.]. On Culture: Suicide notes + social media = troubling trend. http://www.newsobserver.com/living/article15468317.html. Accessed: 2017- 10- 01.
[12] Al Alvarez. 2002. *The savage god: A study of suicide.* A&C Black.
[13] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.. In *LREC*, Vol. 10. 2200–2204.
[14] D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proc. of the International Conference on Learning Representations (ICLR)*. San Diego, CA, USA.
[15] Mark C Baker. 2003. *Lexical categories: Verbs, nouns and adjectives.* Vol. 102. Cambridge University Press.
[16] Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 747–754.
[17] Aaron Ben-Ze'ev. 2004. *Love online: Emotions on the Internet.* Cambridge University Press.
[18] Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions.* Association for Computational Linguistics, 69–72.
[19] Pete Burnap, Walter Colombo, and Jonathan Scourfield. 2015. Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM conference on hypertext & social media.* ACM, 75–84.
[20] Erik Cambria. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems* 31, 2 (2016), 102–107.
[21] Erik Cambria, Soujanya Poria, Rajiv Bajpai, and Björn W Schuller. 2016. SenticNet 4: A Semantic Resource for Sentiment Analysis Based on Conceptual Primitives.. In *COLING*. 2666–2677.
[22] Erik Cambria, Soujanya Poria, Alexander Gelbukh, and Mike Thelwall. 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems* 32, 6 (2017).
[23] Dave Chaffey. 2016. Global Social Media Statistics Summary 2016. *Smart Insights* (2016).
[24] François Chollet et al. 2015. Keras.

[25] Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 51–60.

[26] Glen Coppersmith, Casey Hilland, Ophir Frieder, and Ryan Leary. 2017. Scalable mental health analysis in the clinical whitespace via natural language processing. In *Biomedical & Health Informatics (BHI), 2017 IEEE EMBS International Conference on*. IEEE, 393–396.

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

[28] H Wayland Cummings and Steven L Renshaw. 1979. SLCA III: A metatheoretic approach to the study of language. *Human Communication Research* 5, 4 (1979), 291–300.

[29] Kerstin Denecke and Yihan Deng. 2015. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial intelligence in medicine* 64, 1 (2015), 17–27.

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

[31] Marc Etkind. 1997. *–or Not to be: A Collection of Suicide Notes*. Riverhead Books.

[32] Tom Foster. 2003. Suicide note themes and suicide prevention. *The international Journal of Psychiatry in Medicine* 33, 4 (2003), 323–331.

[33] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. 513–520.

[34] Alex Graves. 2013. Generating Sequences With Recurrent Neural Networks. *CoRR* abs/1308.0850 (2013). http://arxiv.org/abs/1308.0850

[35] Adam Gregory. 1999. The decision to die: The psychology of the suicide note. *Interviewing and deception* (1999), 127–156.

[36] John F Gunn and David Lester. 2015. Twitter postings and suicide: An analysis of the postings of a fatal suicide in the 24 hours prior to death. *Suicidologi* 17, 3 (2015).

[37] Lori D Handelman and David Lester. 2007. The content of suicide notes from attempters and completers. *Crisis* 28, 2 (2007), 102–104.

[38] Marianna E Hayiou-Thomas, Dorothy VM Bishop, and Kim Plunkett. 2004. Simulating SLI: General cognitive processing stressors can produce a specific linguistic profile. *Journal of Speech, Language, and Hearing Research* 47, 6 (2004), 1347–1362.

[39] Kees Hengeveld et al. 1997. Adverbs in Functional Grammar. (1997).

[40] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780.

[41] Maria Ioannou and Agata Debowska. 2014. Genuine and simulated suicide notes: An analysis of content. *Forensic science international* 245 (2014), 151–160.

[42] Mimansa Jaiswal, Sairam Tabibu, and Erik Cambria. 2017. âĂIJHang in ThereâĂİ: Lexical and Visual Analysis to Identify Posts Warranting Empathetic Responses. (2017).

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

[44] Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! *Icwsm* 11, 538-541 (2011), 164.

[45] Antoon Leenaars. 1988. Suicide notes.

[46] David Lester and Antoon A Leenaars. 1988. The moral justification of suicide in suicide notes. *Psychological reports* (1988).

[47] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[48] Saif M Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 shared task on emotion intensity. *arXiv preprint arXiv:1708.03700* (2017).

[49] James W Montgomery. 2000. Relation of working memory to off-line and real-time sentence processing in children with specific language impairment. *Applied Psycholinguistics* 21, 1 (2000), 117–148.

[50] Danielle Mowery, Albert Park, Mike Conway, and Craig Bryan. 2016. Towards automatically classifying depressive symptoms from Twitter data for population health. In *Proceedings of the Workshop on Computational Modeling of PeopleâĂŹs Opinions, Personality, and Emotions in Social Media*. 182–191.

[51] Charles E Osgood. 1960. The cross-cultural generality of visual-verbal synesthetic tendencies. *Systems research and behavioral science* 5, 2 (1960), 146–169.

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

[53] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 79–86.

[54] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.

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

[56] John P Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, K Bretonnel Cohen, John Hurdle, and Christopher Brew. 2012. Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights* 5, Suppl 1 (2012), 3.

[57] Kumar Ravi and Vadlamani Ravi. 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems* 89 (2015), 14–46.

[58] Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 502–518.

[59] Aliza Sarlan, Chayanit Nadam, and Shuib Basri. 2014. Twitter sentiment analysis. In *Information Technology and Multimedia (ICIMU), 2014 International Conference on*. IEEE, 212–216.

[60] Guergana Savova, John Pestian, Brian Connolly, Timothy Miller, Yizhao Ni, and Judith W Dexheimer. 2016. Natural language processing: applications in pediatric research. In *Pediatric biomedical informatics*. Springer, 231–250.

[61] Annika Marie Schoene and Nina Dethlefs. 2016. Automatic Identification of Suicide Notes from Linguistic and Sentiment Features. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. 128–133.

[62] M. Schuster and K.K. Paliwal. 1997. Bidirectional Recurrent Neural Networks. *Trans. Sig. Proc.* 45, 11 (Nov. 1997), 2673–2681.

[63] Jess Jann Shapero. 2011. *The language of suicide notes*. Ph.D. Dissertation. University of Birmingham.

[64] Edwin S Shneidman and Norman L Farberow. 1956. Clues to suicide. *Public health reports* 71, 2 (1956), 109.

[65] Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. 2011. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. 129–136.

[66] Johan AK Suykens and Joos Vandewalle. 1999. Least squares support vector machine classifiers. *Neural processing letters* 9, 3 (1999), 293–300.

[67] Hui Yang, Alistair Willis, Anne De Roeck, and Bashar Nuseibeh. 2012. A hybrid model for automatic emotion recognition in suicide notes. *Biomedical informatics insights* 5, Suppl 1 (2012), 17.

[68] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1480–1489.