Exploring and Learning Suicidal Ideation Connotations on Social Media with Deep Learning

Ramit Sawhney[1], **Prachi Manchanda**[1], **Puneet Mathur**[1],
Raj Singh[1], and **Rajiv Ratn Shah** [4]

[1]Netaji Subhas Institute of Technology, NSIT-Delhi
ramits.co,prachim.co,rajs.co@nsit.net.in, pmathur3k6@gmail.com
[2]Indraprastha Institute of Information Technology, IIIT-Delhi
rajivratn@iiitd.ac.in

Abstract

The increasing suicide rates amongst youth and its high correlation with suicidal ideation expression on social media warrants a deeper investigation into models for the detection of suicidal intent in text such as tweets to enable prevention. However, the complexity of the natural language constructs makes this task very challenging. Deep Learning architectures such as LSTMs, CNNs, and RNNs show promise in sentence level classification problems. This work investigates the ability of deep learning architectures to build an accurate and robust model for suicidal ideation detection and compares their performance with standard baselines in text classification problems. The experimental results reveal the merit in C-LSTM based models as compared to other deep learning and machine learning based classification models for suicidal ideation detection.

1 Introduction

The Centre of Disease Control and Prevention[], in the US, reports that, overall, suicide is the eleventh leading cause of death for all US Americans, and is the third leading cause of death for young people 15-24 years. According to the World Health Organisation(WHO)[], in the last 45 years, suicide rates have increased by 60% worldwide. Suicide attempts are up to 20 times more frequent than completed suicides. Cases of suicide occur due to many complex sociocultural factors and is more likely to occur during periods of socioeconomic, family and individual crisis such as loss of a loved one, unemployment, sexual orientation, difficulties with developing one's identity, disassociation from one's community or other social/belief group, and honour.

The Internet is a powerful source that people turn to when seeking help while having any suicidal thoughts. Understanding how people interact with the Internet in such situations can go a long way in an attempt to prevent such suicides. The availability of suicide-related material on the Internet plays an important role in the process of suicide ideation. Due to this increasing availability of content on social media websites (such as Twitter, Facebook and Reddit etc.), there is an urgent need to identify affected individuals and offer help.

In the past, few attempts such as (O'Dea et al., 2015), (Sueki, 2015) and (Jashinsky et al., 2014) have been made to identify patterns in the language used on social media that express suicidal ideation. However, very few attempts have been made to employ deep learning classifiers that separate text not related to suicide from text that clearly indicates the author exhibiting suicidal intent. In this paper, three deep learning based architectures (Vanilla-RNN, LSTM, and C-LSTM) are compared in the task of sentence classification for the crucial task of detecting suicidal ideation.

The main contributions of this paper can be summarized as follows:

1. Generation of a lexicon of suicide-related words and phrases by scraping suicide web forums to gather tweets for dataset creation.

2. The creation of a labeled dataset for learning the patterns in tweets exhibiting suicidal ideation by manual annotation.

3. Exploration of the performance of three deep learning based architectures for the suicide ideation detection task and compared with

Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 167–175
Brussels, Belgium, October 31, 2018. ©2018 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

three baselines in terms of four evaluation metrics.

2 Related Work

Media communication can have both positive and negative influence on suicidal ideation. A systematic review of all articles in PsycINFO, MEDLINE, EMBASE, Scopus, and CINAH from 1991 to 2011 for language constructs relating to self-harm or suicide by Daine et al. (2013) concluded that internet may be used as an intervention tool for vulnerable individuals under the age of 25. However, not all language constructs containing the word suicide indicate suicidal intent, specific semantic constructs may be used for predicting whether a sentence implies self-harm tendencies or not.

A suicide note analysis method for automated the identification of suicidal ideation was built using binary support vector machine classifiers by Desmet and Hoste (2013) using fine-grained emotion detection for classifier optimization with lexico-semantic features for optimization. In 2014, Huang et al. (2014) used rule-based methods with a hand-crafted unsupervised classification for developing a real-time suicidal ideation detection system deployed over Weibo[1], a microblogging platform. By combining both machine learning and psychological knowledge, they reported an SVM classifier as having the best performance of different classifiers. Some semantic constructs are associated with lifetime suicidal ideation as compared to others. A cross-sectional study of suicidal intent in 220,848 Twitter users in their 20s in Japan (Sueki, 2015) concluded that language framing was important for identifying suicidal markers in the text. For example: *want to suicide* was found to be associated more frequently with a lifetime suicidal intent than *want to die* in similar sentences. Several of these studies emphasized the influencing power of social media and internet in the study of suicide ideation. (Sawhney et al., 2018a) demonstrated the use of ensembles to approach the detection of suicidal mentions on social media.

One of the most concerning issues with suicide-related content on Twitter is the propagation of harmful ideas through social network graphs. A study by Grandjean (2016) performed a classification of users by influence in digital communities based on graph density and vectors of centrality. The study primarily concluded that some users (nodes) in a social network graph had higher influence factor than others. Ueda et al. (2017) collected 1 million tweets following the suicides of 26 prominent figures in Japan between 2010 and 2014 and investigated if media coverage of suicides is correlated with an increase in the actual number of suicides. The reciprocal connectivity between authors of suicidal content suggested a ripple effect in tightly-coupled virtual communities (Colombo et al., 2016) thereby concluding that Twitter is an effective source for investigation of virtual self-harm markers and appropriate intervention. Tweet mining has been successfully been applied in detecting social problems on the web as indicated by Mathur et al. (2018a,b,c) and Mahata et al. (2018).

3 Data

3.1 Data Collection

One of the foremost challenges in the domain of suicidal ideation detection is the lack of availability of a public dataset due to privacy and anonymity concerns borne out of social stigma associated with mental illness and suicide. Motivated by the need to create a fresh dataset, the primary requirement of developing a suicidal language for data collection was identified. Rather than developing a word list to represent this language, a corpus of words and phrases were developed using anonymized data from known Suicide web forums (Burnap et al., 2015). These forums were identified by Recupero et al. (2008) as dedicated for suicidal issues with related discussions in this subject. Between 3rd December 2017 and 31st January 2018, four of these Suicide forums were scraped for the user posts and human annotators were asked to identify if these posts had any suicidal intent. In addition to this, user posts (containing tags of 'suicide') from the micro-blogging websites, Tumblr and Reddit were collected and added to this collection.

This resulted in the following composition of posts: 300 from each of the Suicide forums and 2000 posts randomly selected from the Tumblr and Reddit posts. These were subsequently human annotated based on them having a suicidal intent or not. Then, Term Frequency/Inverse Document frequency (Ramos et al., 2003) (TF-IDF) method was applied to this set of manually annotated texts to identify terms which appear frequently in the

[1] http://www.scmp.com/topics/weibo

suicidal	suicide	not worth living	slit my wrist
kill myself	can’t go on	ready to jump	cut my wrist
my suicide note	want to die	sleep forever	slash my wrist
my suicide letter	be dead	suicide plan	do not want to be here
end my life	better off without me	bold	want it to be over
never wake up	better off dead	bold	want to be dead
suicide pact	don’t want to be here	tired of living	nothing to live for
die alone	go to sleep forever	die now	ready to die
wanna die	wanna suicide	commit suicide	not worth living
why should I continue living	take my own life	thoughts of suicide	I wish I were dead
to take my own life	suicide ideation	depressed	kill me now

Table 1: Words/Phrases linked with Suicidal Intent

Suicidal	Non-suicidal
I want to kill myself	Visit the #SuicideAwarenessCampaign this weekend
I failed again. I can’t do this anymore.	The movie was so bad, I wanted to kill myself.
When did I get addicted? Kill me now!	1 girl commits suicide from EY Square Rooftop
My husband has Cancer. I want to die.	Finish this sentence: Before I die I want to —
My mental illness leaves me only to suicide	An honest talk about the recent suicides in the city.
Suicide is my only really option...	My friend attempted suicide. Weeks later I got this mail.
Life sucks. #gonnasuicide #onthebridge	Idk man. Social media is suicide. Please kill urself

Table 2: Examples of human annotation of tweets

	H_1	H_2	H_3
H_1	–	0.61	0.48
H_2	0.61	–	0.51
H_3	0.48	0.51	–

Table 3: Cohen’s Kappa for three annotators H_1, H_2 and H_3

texts belonging to the suicidal ideation class and less frequently in the non-ideation class. These terms play a role in differentiating between the two classes. Finally, manual annotators were asked to remove any terms from this list which were not based on suicidal intent as well as duplicate terms. This gave a final lexicon of 108 terms consisting of but not limited to the phrases/words of Table 1.

The public Streaming API [2] offered by the microblogging website Twitter allows programmatic collection of tweets as they occur, filtered by specific criteria. Using the same, anonymized data was collected from Twitter. This content contained self-classified suicidal ideation (i.e. text posts tagged or ’hash-tagged’ with a word or phrase present in the generated corpus).

The tweets retrieved from Twitter using the API contain extraneous information. It can be associated with a URL, user mention, media files(image, audio, and video), timestamp, number of retweets. For the tasks in this paper, the text from each tweet was extracted while the rest of the information about the tweet was discarded. Although the tweets were collected from the ’Stream’ based on a suicidal language earlier developed, the exact sentiment of the tweets was unknown. Tweets consisting of suicidal terms could be related to other things as well. Eg. *suicidal awareness campaign and prevention*, *a news report consisting of a third person’s suicide*, *sarcasm* etc. This made a manual annotation of the dataset imperative for better accuracy.

3.2 Data Annotation

The final dataset consisting of 5213 text sentences from different tweets was then, manually annotated. Three human annotators were asked to classify the texts from the dataset based on binary criterion (*Does this text imply self-harm inflicting tendencies or suicidal intent?*). This means that the annotators were asked to select one of the two categories (**Suicidal** or **Non-suicidal**) and to se-

[2] https://developer.twitter.com/en/docs

lect **Suicidal** in case of ambiguity. The suicidal criterion means that the tweet is a clear display of suicidal intent by the user. The suicide is imminent and not conditional unless some event is a clear risk factor eg: depression, bullying, substance abuse. On the other hand, the non-suicidal criteria is the default category for all the texts, i.e. they show no evidence or ambiguous evidence towards suicidal intent. They might include sarcasm, news reports or suicidal awareness texts. The classification is more clearly explained using examples in Table 2.

A satisfactory agreement between the annotators (e.g., 0.51 for H_2 and H_3) can be inferred from Table 3.

As a result, 822 tweets in the dataset (ie., 15.76% of the dataset) were annotated to be suicidal while the rest were classified into ’Non-Suicidal’.

4 Methodology

4.1 Preprocessing

Preprocessing involves filtering the input text to improve the accuracy of the proposed methodology by eliminating redundant features and noise. This is achieved by applying a series of filters, based on Xiang et al. (2012), in the order given below to process the raw tweets prior to learning the word embeddings.

1. Removal of non-English tweets using LingPipe (Baldwin and Carpenter, 2003) with Hadoop.

2. Removal of URLs in tweets.

3. Identification and elimination of user mentions in tweet bodies having the format of @username as well as retweets in the format of RT.

4. Removal of all hashtags with length $>$ 10 due to a great volume of hashtags being concatenated words, which tends to amplify the vocabulary size inadvertently and leads to redundant features.

5. Condensation of three or more than three repetitive letters into a single letter, e.g. *dieeee* to *die*. Similar heuristics have been used in other work such as (Go et al., 2009).

6. Stopword removal.

7. Removal of tokens that are not a sequence of letters, - or ’. This includes removal of numbers, terms such as *h31100oo*, etc, which do not represent words.

4.2 Distributed Word Representation

A distributed language representation X consists of an embedding for every vocabulary word in space S with dimension D, the dimension of the latent representation space. The embeddings are learned to optimize an objective function defined on the original text, such as the likelihood of word occurrences. An interesting implementation to get the word embeddings is the word2vec model (Mikolov et al., 2013a) which is used here.

word2vec is a group of related models that are used to produce word embeddings. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space.

Generating word embeddings from text corpus is an unsupervised process. To get high-quality embedding vectors, a large amount of training data is necessary. After training, each word, including all hashtags, is represented by a real-valued vector which can be given as input to a deep learning based model.

4.3 Deep Learning Models

An efficient model to classify sequential information of arbitrary length is a Recurrent Neural Network(RNN) (Elman, 1990) model.

However, the gradient vector of RNNs with transition functions of this form can grow or decay exponentially over long sequences (Hochreiter et al., 2001) which makes it difficult to learn long distance correlations.

Long Short Term Memory (Hochreiter and Schmidhuber, 1997) prevents this vanishing or explosion gradient seen in the RNN and is thus, preferred over RNN. The LSTM has a memory cell which consists of four main components: input, output, forget gates and candidate memory cell. The forget gates control the information that

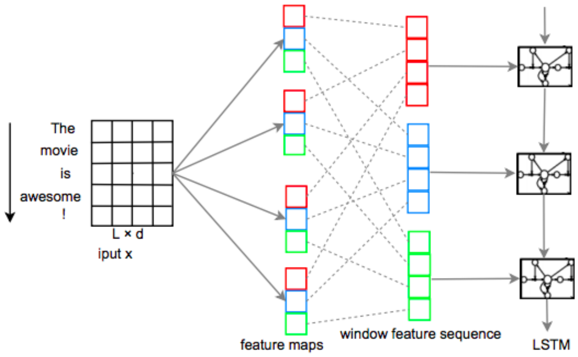

Figure 1: The architecture of C-LSTM for sentence modeling taken from (Zhou et al., 2015). Blocks of the same color in the feature map layer and window feature sequence layer corresponds to features for the same window. The dashed lines connect the feature of a window with the source feature map. The final output of the entire model is the last hidden unit of LSTM.

is to be sent to the next time step. The memory cell stores the data at each step and thus ensures long-distance correlations. The output at each time step depends on the input of that step, the output from the previous time step, the forget gates and the data in the memory cell.

The LSTM architecture is similar to a standard RNN. At each time step, the output of the module is controlled by a set of gates in R^d as a function of the old hidden state h_{t1} and the input at the current time step x_t: the forget gate f_t, the input gate i_t and the output gate o_t. These gates collectively decide how to update the current memory cell C_t and the current hidden state h_t. We use d to denote the memory dimension in the LSTM and all vectors in this architecture share the same dimension.

LSTMs are well-suited to classify, process and predict time series and capture long-term dependencies in sentences along with a relative insensitivity to gap length unlike alternative models such as RNNs and hidden Markov Models (Eddy, 1996) make it an excellent choice for the identification of suicidal ideation in tweets.

In a C-LSTM Model (Zhou et al., 2015), CNN and LSTM are stacked in a semantic sentence modelling. As is shown in Figure 1, the CNN is applied to text data and consecutive window features which are extracted are fed into the LSTM model which enables it to learn long-range dependencies from higher-order sequential features. The one-dimensional convolution involves a filter vector sliding over a sequence and detecting features at different positions. The C-LSTM model uses multiple filters to generate multiple feature maps which are rearranged as feature representations for each window. The new successive higher-order window representations then are fed into LSTM. The output of the hidden state at the last time step of the LSTM is regarded as the document representation. The efficient spatial encoding and automatic feature extraction by the CNN layer combined with the efficient text classification by LSTMs motivate this study to explore the C-LSTM model for suicidal ideation identification.

4.4 Classification

Suicidal Ideation detection is formulated as a supervised binary classification problem. For every tweet $t_i \in D$, the dataset, a binary valued variable $y_i \in \{0, 1\}$ is introduced, where $y_i = 1$ denotes that the tweet t_i exhibits Suicidal Ideation. To learn this, the classifier must determine whether any sentence in t_i possesses a certain structure or keywords that mark the existence of any possible Suicidal thoughts. The word embeddings derived from the previous step are used to train a classification model to identify tweets exhibiting suicidal ideation. Three Deep Learning based architectures, namely, vanilla RNN, vanilla LSTM and C-LSTM, are explored for the suicidal ideation detection task. The architectural is presented in the following section.

The following steps are executed on every tweet $t_i \in D$:

1. *Word Embeddings.* Top-N frequent words occurring in a tweet are encoded to form an embedding layer utilizing the 300-dimensional word2vec embeddings.

2. *Sentence Embeddings.* For the C-LSTM model, a one dimensional CNN and max-pooling layer are added after the embedding layer. These sentence embeddings are then fed into the LSTM layer.

3. *Classification.* Ultimately, the model feeds the learned sentence embeddings *(C-LSTM)* or word embeddings *(Vanilla RNN or LSTM)* to a deep neural network *(RNN or LSTM).*

Model	Accuracy	Precision	Recall	F1 Score
LR: Character n-grams	0.669	0.663	0.753	0.702
LR: TF-IDF	0.727	0.767	0.778	0.772
LR: Bag of Words	0.737	0.712	0.788	0.748
SVM: Character n-grams	0.682	0.676	0.763	0.713
SVM: TF-IDF	0.732	0.724	0.793	0.758
SVM: Bag of Words	0.730	0.712	0.801	0.733
RNN	0.737	0.720	0.817	0.753
LSTM	0.789	0.745	**0.874**	0.796
C-LSTM	**0.812**	**0.787**	0.872	**0.827**

Table 4: Classification Results in terms of Evaluation metrics.

5 Experiment Settings

5.1 Baselines

In order to offer fair comparisons to other competitive models, and validate the proposed Deep Learning model, experiments are conducted with baselines. Hand-crafted features are extracted from tweets and are fed into a linear classifier. Multinomial logistic regression (Böhning, 1992) is used as a classifier with the three feature extraction models given below. Support Vector Machines have also been used for feature extraction based suicide-ideation classification problems, and hence are also used as baselines to compare performance. 10-fold cross-validation is performed to report results in terms of the evaluation metrics presented in the following subsections.

1. *Character n-grams*. State-of-the-art method (Cavnar et al., 1994) for sentence level classification using up to 3-grams from each tweet.

2. *TF-IDF*. Text Frequency - Inverse Document Frequency (TF-IDF) are commonly used features for text classification.

3. *Bag of Words*. A bag-of-words model (Sriram et al., 2010) is constructed by selecting the 50,000 most frequent words from the training tweets. The count of each word is used to create a feature vector for classification.

5.2 Model Architectures and Parameters

For the classification task, both a RNN and a LSTM are trained using 10-fold cross-validation to identify the best hyper-parameter settings. Pre-Trained word2vec word embeddings that were trained on 100 billion words from Google News are employed as features for classification. These vectors have a dimensionality of 300 and were trained using the continuous bag-of-words architecture (Mikolov et al., 2013b). The experiment settings pertaining to both are presented below:

1. *RNN*. Vanilla RNN with $h = 128$ units, 32 dense units, a dropout rate of 0.1.

2. *LSTM*. Vanilla LSTM with $h = 128$ memory units, 32 dense units, a dropout probability of 0.2.

3. *C-LSTM*. Convolution Layer (mask size $=$ 5, filter maps$=$ 128)$\rightarrow$ Max-Pooling Layer (mask size $=$ 2) $\rightarrow$ LSTM layer (h $=$ 128) $\rightarrow$ Dropout Layer with dropout probability = 0.2.

ReLU (Nair and Hinton, 2010) was used for activation the CNN layers in C-LSTM, and Dense layer with single neuron and sigmoid activation was used for all the models. Dropout layers were added to all models to avoid over-fitting. A batch size of 64 was chosen, and the models were trained for a total of 10 epochs. The Adam Optimizer (Kingma and Ba, 2014) was used to minimize log loss.

5.3 Evaluation Metrics

The Baselines and Deep learning models above are compared with each other in terms of the following metrics:

1. **Precision** $= \frac{t_p}{t_p+f_p}$

2. **Recall** $= \frac{t_p}{t_p+f_n}$

3. **F1 score** $= \frac{2t_p}{2t_p+f_p+f_n}$

4. **Accuracy** $= \frac{t_p+t_n}{t_p+t_n+f_p+f_n}$

where, t_p is the number of true positives, t_n is the number of true negatives, f_p is the number of false positives, and f_n is the number of false negatives.

5.4 Results and Analysis

Table 4 shows the results of the baselines as well as deep learning models on the suicide ideation detection task in terms of the evaluation metrics. The first six rows show results for baseline methods, whereas the bottom three rows focus on proposed deep learning models. The results shown are obtained using 10-fold cross validation.

As the table shows, C-LSTM perform performs significantly better than the baseline methods as well as vanilla LSTM and RNN. This is attributed to the ability of LSTMs to learn how to forget past observations makes them more robust to noise, and better able to capture long-term dependencies in a tweet combined with the efficient encoding of the one-dimensional spatial structure in the sequence of words for tweets which further serve as input to the LSTM layer. RNNs are comparable to both TF-IDF and Bag of words models with Multinomial logistic Regression and Support Vector Machines. Among the baselines, the TF-IDF model combined with multinomial logistic regression is better than the others. Surprisingly, standard feature extraction methods coupled with a linear classifier perform comparatively well as compared to RNNs that involve a much larger amount of computation. However, there is a vast improvement with the incorporation of LSTM and C-LSTM which easily compensates for the additional computation involved.

5.5 Error Analysis

A brief error analysis is presented in this subsection to highlight some of the tweets both annotators and the proposed models that gave erroneous results.

- **Subtle references** *Life is so meaningless to me right now, should prolly end it* The models were unable to identify the subtle hints towards suicidal ideation.

- **Uncertainty** *Friends are worrying about me committing suicide.* It is unclear for both annotators and the system to identify the nature of this tweet due to the lack of explicit suicidal intent.

- **Unfamiliarity** *I finally found a whole bottle full of pills, im sorry* The current training dataset lacks in terms of suicidal ideation phrases and would need updates to cover broader aspects and learn the context between topics such as pill overdose and suicides.

6 Conclusion and Future Work

In this paper, three Deep Learning based models, particularly RNN, LSTM, and C-LSTM are employed for the task of suicidal ideation detection in tweets. For this purpose, a lexicon of terms was first generated by scraping and manually annotating anonymized data from known suicide Web forums. A dataset of tweets was collected using the Twitter REST API by using search queries corresponding to the generated lexicon. Human annotators labeled tweets with suicidal intent present or absent, which were then used to train both three machine learning-based baseline models as well as the three proposed deep learning models. A quantitative comparison between the various models revealed the effectiveness of a C-LSTM based model in suicidal ideation detection in tweets. This was attributed to the ability of CNNs to spatially encode the tweets into a one-dimensional structure to be fed into LSTMs along with the ability of LSTMs to capture long-term dependencies. In the future, this work can be extended by investigating other deep learning based architectures for the tasks of suicidal ideation detection on Twitter as well as other Web forums and Social media. Also, nature-inspired heuristics can be explored for efficient feature selection as done by Sawhney et al. (2018b,c).

References

Breck Baldwin and Bob Carpenter. 2003. Lingpipe. *Available from World Wide Web: http://alias-i.com/lingpipe*.

Dankmar Böhning. 1992. Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics*, 44(1):197–200.

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*, pages 75–84. ACM.

William B Cavnar, John M Trenkle, et al. 1994. N-gram-based text categorization. *Ann arbor mi*, 48113(2):161–175.

Gualtiero B Colombo, Pete Burnap, Andrei Hodorog, and Jonathan Scourfield. 2016. Analysing the connectivity and communication of suicidal users on twitter. *Computer communications*, 73:291–300.

Kate Daine, Keith Hawton, Vinod Singaravelu, Anne Stewart, Sue Simkin, and Paul Montgomery. 2013. The power of the web: a systematic review of studies of the influence of the internet on self-harm and suicide in young people. *PloS one*, 8(10):e77555.

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

Sean R Eddy. 1996. Hidden markov models. *Current opinion in structural biology*, 6(3):361–365.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12).

Martin Grandjean. 2016. A social network analysis of twitter: Mapping the digital humanities community. *Cogent Arts & Humanities*, 3(1):1171458.

Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, et al. 2001. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.

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

Xiaolei Huang, Lei Zhang, David Chiu, Tianli Liu, Xin Li, and Tingshao Zhu. 2014. Detecting suicidal ideation in chinese microblogs with psychological lexicons. In *Ubiquitous Intelligence and Computing, 2014 IEEE 11th Intl Conf on and IEEE 11th Intl Conf on and Autonomic and Trusted Computing, and IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UTC-ATC-ScalCom)*, pages 844–849. IEEE.

Jared Jashinsky, Scott H Burton, Carl L Hanson, Josh West, Christophe Giraud-Carrier, Michael D Barnes, and Trenton Argyle. 2014. Tracking suicide risk factors through twitter in the us. *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, 35(1):51.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Debanjan Mahata, Jasper Friedrichs, Rajiv Ratn Shah, and Jing Jiang. 2018. Did you take the pill?-detecting personal intake of medicine from twitter. *arXiv preprint arXiv:1808.02082*.

Puneet Mathur, Meghna Ayyar, Sahil Chopra, Simra Shahid, Laiba Mehnaz, and Rajiv Shah. 2018a. Identification of emergency blood donation request on twitter. In *Proceedings of the Third Workshop On Social Media Mining for Health Applications*.

Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. 2018b. Did you offend me? classification of offensive tweets in hinglish language. In *Proceedings of the Second Workshop on Abusive Language Online*.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018c. Detecting offensive tweets in hindi-english code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Bridianne O'Dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on twitter. *Internet Interventions*, 2(2):183–188.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142.

Patricia R Recupero, Samara E Harms, and Jeffrey M Noble. 2008. Googling suicide: surfing for suicide information on the internet. *The Journal of clinical psychiatry*.

Ramit Sawhney, Prachi Manchanda, Raj Singh, and Swati Aggarwal. 2018a. A computational approach to feature extraction for identification of suicidal ideation in tweets. In *Proceedings of ACL 2018, Student Research Workshop*, pages 91–98.

Ramit Sawhney, Puneet Mathur, and Ravi Shankar. 2018b. A firefly algorithm based wrapper-penalty feature selection method for cancer diagnosis. In *International Conference on Computational Science and Its Applications*, pages 438–449. Springer.

Ramit Sawhney, Ravi Shankar, and Roopal Jain. 2018c. A comparative study of transfer functions in binary evolutionary algorithms for single objective optimization. In *International Symposium on Distributed Computing and Artificial Intelligence*, pages 27–35. Springer.

Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. 2010. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 841–842. ACM.

Hajime Sueki. 2015. The association of suicide-related twitter use with suicidal behaviour: a cross-sectional study of young internet users in japan. *Journal of affective disorders*, 170:155–160.

Michiko Ueda, Kota Mori, Tetsuya Matsubayashi, and Yasuyuki Sawada. 2017. Tweeting celebrity suicides: Users' reaction to prominent suicide deaths on twitter and subsequent increases in actual suicides. *Social Science & Medicine*, 189:158–166.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1980–1984. ACM.

Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis C. M. Lau. 2015. A C-LSTM neural network for text classification. *CoRR*, abs/1511.08630.