Learning Models for Suicide Prediction from Social Media Posts[*]

Ning Wang[1] **Fan Luo**[1] **Yuvraj Shivtare**[1†] **Varsha D. Badal**[2]
K.P. Subbalakshmi[1] **R. Chandramouli**[1] **Ellen Lee**[2]
Stevens Institute of Technology[1]
Hoboken, NJ 07030
{nwang7, fluo4, yshivtar, ksubbala, mouli}@stevens.edu
University of California San Diego[2]
San Diego, CA 92161
{vbadal, eel013}@health.ucsd.edu

Abstract

We propose a deep learning architecture and test three other machine learning models to automatically detect individuals that will attempt suicide within (1) 30 days and (2) six months, using their social media post data provided in (Macavaney et al., 2021) via the CLPsych 2021 shared task. Additionally, we create and extract three sets of handcrafted features for suicide risk detection based on the three-stage theory of suicide and prior work on emotions and the use of pronouns among persons exhibiting suicidal ideations. Extensive experimentations show that some of the traditional machine learning methods outperform the baseline with an F1 score of 0.741 and F2 score of 0.833 on subtask 1 (prediction of a suicide attempt 30 days prior). However, the proposed deep learning method outperforms the baseline with F1 score of 0.737 and F2 score of 0.843 on subtask 2 (prediction of suicide 6 months prior).

1 Introduction

According to World Health Organization (WHO) [1], close to 800,000 people die due to suicide every year, which is one person every 40 seconds. The US Centers for Disease Control and Prevention (CDC) [2] claimed that suicide was the tenth leading cause of death overall in the United States. Recently, there has been a trend in using natural language processing (NLP) techniques on unstructured physician notes from electronic health record (EHR) data to detect high-risk patients (Fernandes et al., 2018).

With the proliferation of social media where there is free sharing of information, mining data from these platforms has become a natural way to extend the above body of work in more natural settings. Consequently, researchers have started to apply machine learning and NLP based techniques to detect suicide ideation on social media platforms (Ramírez-Cifuentes et al., 2020; Roy et al., 2020). Some of them focused on handcrafted features, including TF-IDF (Zhang et al., 2011), LIWC (Tausczik and W, 2010), N-gram, Part-of-Speech (PoS) and emotions (Shah et al., 2020; Zirikly et al., 2019; Zhang et al., 2015; Ji et al., 2020), while others explored language embeddings (Cao et al., 2019; Jones et al., 2019; Sawhney et al., 2018; Coppersmith et al., 2018).

In this paper, we present several approaches to detect suicide ideation from Twitter posts (1) 30 days before the attempt and (2) six months before the attempt. We use the dataset provided by the CLPsych 2021 Shared Tasks Macavaney et al. (2021) towards this goal.

The main contributions of our work are:

- Explored and generated multiple handcrafted feature sets motivated by prior work in this area
- Proposed a new deep learning architecture that uses latent features from tweets to detect suicide attempts
- Tested several machine learning algorithms using only handcrafted features and only latent features
- Achieved better performance than baseline in terms of F1, F2 and True Positive Rate (TPR) on both subtasks

Summary of Findings: The main takeaways from this work are:

- Extensive testing on the dataset shows that latent feature (Doc2Vec (Lau and Baldwin, 2016)), is better at detecting suicide attempts from the tweets than handcrafted features

[*] This work is accepted to CLPsych 2021, to be held in conjunction with NAACL 2021

[†] Equal contribution to 2nd author

[1] https://www.who.int/

[2] https://www.cdc.gov/

- Most of our models performed better on detecting individuals who have attempted suicide or were a victim of suicide than on detecting control individuals who have not
- The KNN and SVM with latent features perform best on subtask 1, with respect to F1, F2 and TPR; while our proposed C-Attention (C-Att) network performs best on subtask 2, with respect to F1, F2 and TPR

2 Method

Before we describe the methods in detail we provide a summary of the features used in our work. We use two classes of features: latent features and handcrafted features. These are described in the sections below.

2.1 Latent Features

Latent features are typically obtained as language embeddings. In our case, we used the Doc2vec (Lau and Baldwin, 2016) to generate both word embeddings and document embeddings on each post. Doc2Vec creates a vectorized representation of a group of words (or a single word, when used in that mode) taken collectively as a single unit. For every document in the corpus, Doc2Vec computes a feature vector. There are two models for implementing Doc2vec: Distributed Memory version of Paragraph Vector (PV-DM) and Distributed Bag of Words version of Paragraph Vector (PV-DBOW). For our experimentation, we used Distributed Memory (DM) version. DM randomly samples consecutive words from a sentence and predicts a center word using these randomly sampled set of context words and the feature vector.

2.2 Handcrafted Features

2.2.1 Emotions

Emotions can be good indicators of depression and suicide ideation (Desmet and Hoste, 2013; Coppersmith et al., 2016; Cao et al., 2020; Ghosh et al., 2020), so we include emotions as one of the handcrafted features. We used the method proposed in (Shao et al., 2019) to generate 12 emotion tags, including contentment, pride, fear, anxiety, sadness, disgust, relief, shame, anger, interest, agreeableness and joy. Apart from that we also generated emotion intensity scores using NRC lexicon (Mohammad, 2018), for the emotions like anger, anticipation, disgust, fear, joy, sadness, surprise and trust. After removing duplicates, we selected 17 emotion tags.

2.2.2 Parts of Speech

We use NLTK (Bird et al., 2009) to generate Part-of-Speech tags. PoS tags can detect the syntactic structure difference between users that attempt suicide and the control group (Ji et al., 2020). It has been shown (Roubidoux, 2012) that persons attempting suicide use more first person pronouns. Therefore, we also calculate the number of occurrences of first person pronouns like “I", “me", “mine" and “myself" and include this count as another PoS related handcrafted feature. In total, we generated 34 PoS tags per post for the “30 days prior prediction" subtask and 37 PoS tags for the “6 months prior prediction" subtask.

2.2.3 Three-step theory of suicide and suicide dictionary

We then generate a dictionary of words based on the three-step theory of suicide (3ST) (Klonsky and May, 2015) beginning with the ideation, followed by unmitigated strengthening of the idea due to insufficient social support and precipitated by an attempt. These stages are underpinned by feelings of hopelessness (Dixon et al., 1991), thwarted belongingness and burdensomeness (Chu et al., 2018; Forkmann and Teismann, 2017). Violence usually differentiates attempters and non-attempters (Stack, 2014). Surviving an attempt is expected to be accompanied by feelings of shame (Wiklander et al., 2012; Wolk-Wasserman, 1985). We expect these feelings to be out of phase with each other creating a leading, inline and lagging indicator of suicide attempt. We used Word2vec (Mikolov et al., 2013b,a,c) software to construct these dictionaries using the accompanying utility (also available in online versions) by evaluating closest neighbors of words (gloom and burden, violence, hurt and shame), each containing about 100 words with some manual cleanup and editing. The manual cleanup involved removing stop-words, words with hyphens, special characters, some vernacular tokens, and words that differed in capitalization alone. We generated this feature set by counting each keyword in each post. In addition, we manually created a dictionary of suicide keywords based on suicide-related words published in (Low et al., 2020; Yao et al., 2020), and counted how many suicide-related keywords occurred in each post. [3]

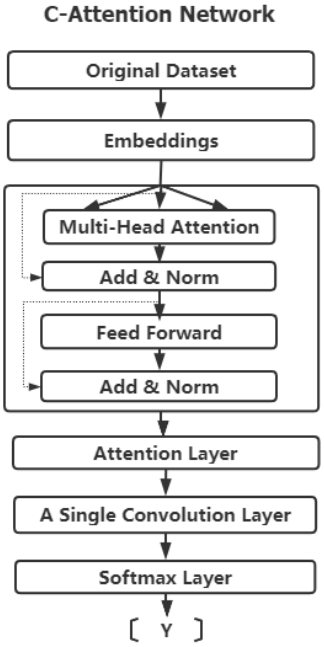

Figure 1: The proposed architecture of C-Attention Network

2.3 Models

In this work, we proposed a deep learning model and used a few other machine learning models for each subtask. The proposed deep learning model, which we refer to as the C-Attention Network, is our primary model.

2.3.1 C-Attention Network

Figure 1 depicts our C-Attention network which uses latent features to detect suicide attempts. This network is similar to our prior C-Attention Embedding model (Wang et al., 2020) with the following differences:

- In this work we consider each post as a small document, and use Doc2Vec to generate a 100-dimension embedding representation for each post; whereas the work in (Wang et al., 2020) generated a sentence embedding for each sentence in a speech.
- We removed the positional encoding layer since there is no positional dependency among posts.

In summary, the architecture first calculates the embeddings of the dataset, then processes it via a multi head self-attention (MHA) module that captures the intra-feature relation-ships; an attention layer followed by a single convolution layer and a softmax layer. The MHA module is the same as that proposed in (Vaswani et al., 2017) for the popular transformer architecture.

2.3.2 Latent Features with Other Machine Learning Models

In this approach we combined all the posts for each user. Stop words were removed from the posts and lemmatized. The average length of posts was found to be 140 words. Long posts were chunked into 150 words segments to retain meaningful information in each post. A single 200-dimension embedding vector is generated for each segment using the Doc2Vec as described in Section 2.1.

We applied linear discriminant analysis (LDA) (McLachlan, 2004) for dimensionality reduction before classification. The output of LDA was fed to machine learning models. K-Nearest Neighbors (KNN) (Jiang et al., 2012) with K=3, Support Vector Machine (SVM) (Ríssola et al., 2019) with linear kernel (referred to as SVM(EB) in the rest of the paper) and Decision Tree (D-Tree) (Song and Ying, 2015) classifier models were considered.

2.3.3 Handcrafted Features with Other Machine Learning Models

We used three other machine learning models on the handcrafted features described in Sec 2.2 to address both challenges. The three machine learning models were: Random Forest Classifier (RF) (Breiman, 2001), Logistic Regression (LR) (Aladağ et al., 2018) and Support Vector Machine (SVM) (Ríssola et al., 2019) (referred to as SVM(HF) for the rest of the paper). We used the entire handcrafted features since we found that leaving out any of those handcrafted feature sets would introduce a performance drop. We fine-tuned the parameters of each ML model, for example, we set the kernel as rbf (radial basis function) on SVM(HF) model; set the solver as liblinear (limited to one-versus-rest schemes) on LR model; and set the max depth to 4 on RF model to get the best predictions.

[3] Available at: https://sites.google.com/stevens.edu/infinitylab/suicide-risk-detection

	F1	F2	TPR	FPR	AUC
Subtask 1 (30 days)					
Baseline	0.636	0.636	0.636	0.364	0.661
KNN	0.286	0.278	0.273	0.636	0.264
SVM(EB)	0.400	0.377	0.364	0.455	0.529
SVM(HF)	0.364	0.364	0.364	0.636	0.397
Subtask 2 (6 months)					
Baseline	0.710	0.724	0.733	0.333	0.764
KNN	0.429	0.411	0.400	0.467	0.444
SVM(EB)	0.533	0.533	0.533	0.467	0.640
SVM(HF)	0.400	0.400	0.400	0.600	0.502

Table 1: Results obtained by running the KNN, SVM(EB) and SVM(HF) models trained on the entire training set.

3 Results

Table 1 and Table 2 show the performance results. The results reported in Table 1 were obtained by running the KNN, SVM(EB) and SVM(HF) models which were trained on the entire training set. The performance of the models are measured in terms of F1 and F2 scores, True Positive Rates (TPR), False Positive Rates (FPR) and Area Under the ROC Curve (AUC).

4 Analysis/Discussion

The results reported in Table 1 were generated by the KNN, SVM(EB) and SVM(HF) models, which performed best on the training set. From Table 1, we can see that the baseline provided by the CLPsych 2021 shared task outperformed all of these methods. After a thorough investigation of the results, we observed that those models that did not perform best on the training set, performed better on the test set. It probably indicates that we over-trained our models on the training set.

As a result, in the following experiments, we randomly split the training set into 80% for training and 20% for validation, and use the models that performed best on the validation set to predict suicide in the test set. The new performance results on the test set are shown in Table 2.

We noted that in subtask 1, KNN and SVM(EB) performed best in terms of F1, F2 and TRP. The best AUC was achieved by KNN only, and the best FPR was achieved by RF. In subtask 2, C-Att performed best in terms of F1, F2 and TRP; the best FPR was achieved by RF; and the best AUC was achieved by Baseline.

Our experiment results would indicate that:

	F1	F2	TPR	FPR	AUC
Subtask 1 (30 days)					
Baseline	0.636	0.636	0.636	0.364	0.661
C-Att	0.690	0.806	**0.909**	0.727	0.504
SVM(HF)	0.621	0.726	0.818	0.818	0.570
LR	0.571	0.556	0.545	0.364	0.434
RF	0.444	0.392	0.364	**0.273**	0.603
KNN	**0.741**	**0.833**	**0.909**	0.545	**0.694**
D-Tree	0.667	0.750	0.818	0.636	0.591
SVM(EB)	**0.741**	**0.833**	**0.909**	0.545	0.653
Subtask 2 (6 months)					
Baseline	0.710	0.724	0.733	0.333	**0.764**
C-Att	**0.737**	**0.843**	**0.933**	0.600	0.76
SVM(HF)	0.600	0.706	0.800	0.867	0.518
LR	0.563	0.584	0.600	0.533	0.542
RF	0.417	0.362	0.333	**0.267**	0.558
KNN	0.500	0.479	0.467	0.400	0.536
D-Tree	0.500	0.479	0.467	0.400	0.533
SVM(EB)	0.444	0.417	0.400	0.400	0.489

Table 2: Results obtained when the training dataset was split into training and validation set as described. HF represents handcrafted features. EB represents word embeddings.

- In general, latent features perform better than handcrafted features in this shared task
- C-Att model performs better on longer range suicide predictions and KNN and SVM(EB) work better on shorter range suicide predictions
- Besides RF, our other models perform better on detecting suicide individuals than control individuals

5 Conclusion

In this work, we introduce C-Attention model and test other machine learning models to automatically detect suicidal individuals based on the latent feature (Doc2Vec) and handcrafted features including emotions, PoS, and three-step theory of suicide and suicide dictionary. Our results show that both KNN and SVM(EB) achieved the best F1 score of 0.741 and F2 score of 0.833 on subtask 1 (prediction of a suicide attempt 30 days prior), and C-Att reached the best F1 score of 0.737 and F2 score of 0.843 on subtask 2 (prediction of suicide 6 months prior).

Ultimately, this work supports the use of social media as an avenue to better predict and understand the experience of suicidal thoughts. However more work is needed to better decipher why certain features and models best predict suicidality in large, diverse, representative samples.

Ethics Statement

Secure access to the shared task dataset was provided with IRB approval under University of Maryland, College Park protocol 1642625.

Acknowledgements

We appreciate the efforts of the organizers of this challenge to make the data and computational resources available to us.

The organizers are particularly grateful to the users who donated data to the OurDataHelps project without whom this work would not be possible, to Qntfy for supporting the OurDataHelps project and making the data available, to NORC for creating and administering the secure infrastructure, and to Amazon for supporting this research with computational resources on AWS.

References

Ahmet Emre Aladağ, Serra Muderrisoglu, Naz Berfu Akbas, Oguzhan Zahmacioglu, and Haluk O Bingol. 2018. Detecting suicidal ideation on forums: Proof-of-concept study. *JMIR*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Leo Breiman. 2001. Random forests. *Springer*.

Lei Cao, Huijun Zhang, and Ling Feng. 2020. Building and using personal knowledge graph to improve suicidal ideation detection on social media. *IEEE*.

Lei Cao, Huijun Zhang, Ling Feng, Zihan Wei, Xin Wang, Ningyun Li, and He Xiaohao. 2019. Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. *ArXiv*.

Carol Chu, Megan L Rogers, Anna R Gai, and Thomas E Joiner. 2018. Role of thwarted belongingness and perceived burdensomeness in the relationship between violent daydreaming and suicidal ideation in two adult samples. *Journal of aggression, conflict and peace research*.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the third workshop on computational linguistics and clinical psychology*, pages 106–117.

Bart Desmet and Véronique Hoste. 2013. Emotion detection in suicide notes. *ScienceDirect*.

Wayne A Dixon, P Paul Heppner, and Wayne P Anderson. 1991. Problem-solving appraisal, stress, hopelessness, and suicide ideation in a college population. *Journal of Counseling Psychology*, 38(1):51.

Andrea C Fernandes, Rina Dutta, Sumithra Velupillai, Jyoti Sanyal, Robert Stewart, and David Chandran. 2018. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. *Scientific reports*, 8(1):1–10.

Thomas Forkmann and Tobias Teismann. 2017. Entrapment, perceived burdensomeness and thwarted belongingness as predictors of suicide ideation. *Psychiatry research*, 257:84–86.

Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes. *Springer*.

Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. 2020. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE*.

Shengyi Jiang, Guansong Pang, Meiling Wu, and Limin Kuang. 2012. An improved k-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications*, 39(1):1503–1509.

Noah Jones, Natasha Jaques, Pat Pataranutaporn, Asma Ghandeharioun, and Picard Rosalind. 2019. Analysis of online suicide risk with document embeddings and latent dirichlet allocation. *IEEE*.

E David Klonsky and Alexis M May. 2015. The three-step theory (3st): A new theory of suicide rooted in the "ideation-to-action" framework. *International Journal of Cognitive Therapy*, 8(2):114–129.

Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368*.

Daniel M Low, Laurie Rumker, Tanya Talkar, John Torous, Guillermo Cecchi, and Satrajit S Ghosh. 2020. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of medical Internet research*, 22(10):e22635.

Sean Macavaney, Anjali Mittu, Glen Coppersmith, Jeff Leintz, and Philip Resnik. 2021. Community-level research on suicidality prediction in a secure environment: Overview of the CLPsych 2021 shared task. In *Proceedings of the Seventh Workshop on*

Computational Linguistics and Clinical Psychology (CLPsych 2021). Association for Computational Linguistics.

Geoffrey J McLachlan. 2004. *Discriminant analysis and statistical pattern recognition*, volume 544. John Wiley & Sons.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013c. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Saif M. Mohammad. 2018. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

Diana Ramírez-Cifuentes, Freire Ana, Ricardo Baeza-Yates, Joaquim Puntí, Pilar Medina-Bravo, Diego Alejandro Velazquez, Josep Maria Gonfaus, and Jordi Gonzàlez. 2020. Detection of suicidal ideation on social media: multimodal, relational, and behavioral analysis. 22(7):e17758.

Susan M. Roubidoux. 2012. Linguistic manifestations of power in suicide notes: An investigation of personal pronouns.

Arunima Roy, Katerina Nikolitch, Rachel McGinn, Safiya Jinah, William Klement, and Zachary A Kaminsky. 2020. A machine learning approach predicts future risk to suicidal ideation from social media data. *NPJ digital medicine*, 3(1):1–12.

Esteban Ríssola, Diana Ramírez-Cifuentes, Ana Freire, and Fabio Crestani. 2019. Suicide risk assessment on social media: Usi-upf at the clpsych 2019 shared task. *ACL*.

Ramit Sawhney, Prachi Manchanda, Puneet Mathur, Raj Singh, and Shah Rajiv Ratn. 2018. Exploring and learning suicidal ideation connotations on social media with deep learning. *ACL*.

Faisal Muhammad Shah, Farsheed Haque, Ragib Un Nur, Shaeekh Al Jahan, and Zarar Mamud. 2020. A hybridized feature extraction approach to suicidal ideation detection from social media post. *IEEE*.

Zongru Shao, Rajarathnam Chandramouli, KP Subbalakshmi, and Constantine T Boyadjiev. 2019. An analytical system for user emotion extraction, mental state modeling, and rating. *Expert Systems with Applications*, 124:82–96.

Yan-Yan Song and LU Ying. 2015. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130.

Steven Stack. 2014. Differentiating suicide ideators from attempters: Violence—a research note. *Suicide and Life-Threatening Behavior*, 44(1):46–57.

Yla R Tausczik and Pennebaker James W. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *SAGE*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Ning Wang, Mingxuan Chen, and Koduvayur P Subbalakshmi. 2020. Explainable CNN-attention networks (c-attention network) for automated detection of alzheimer's disease. ACM SIGKDD.

Maria Wiklander, Mats Samuelsson, Jussi Jokinen, Åsa Nilsonne, Alexander Wilczek, Gunnar Rylander, and Marie Åsberg. 2012. Shame-proneness in attempted suicide patients. *BMC psychiatry*, 12(1):1–9.

Danuta Wolk-Wasserman. 1985. The intensive care unit and the suicide attempt patient. *Acta Psychiatrica Scandinavica*, 71(6):581–595.

Hannah Yao, Sina Rashidian, Xinyu Dong, Hongyi Duanmu, Richard N Rosenthal, and Fusheng Wang. 2020. Detection of suicidality among opioid users on reddit: Machine learning–based approach. *Journal of medical internet research*, 22(11):e15293.

Lei Zhang, Xiaolei Huang, Tianli Liu, Ang Li, Zhenxiang Chen, and Zhu Tingshao. 2015. Using linguistic features to estimate suicide probability of chinese microblog users. *Springer*.

Wen Zhang, Taketoshi Yoshida, and Xijin Tang. 2011. A comparative study of tf*idf, lsi and multi-words for text classification. *ScienceDirect*.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. *ACL*.