Detecting Suicidality with a Contextual Graph Neural Network

Daeun Lee, Migyeong Kang, Minji Kim, Jinyoung Han*
Sungkyunkwan University
{delee12, gy77, m5512m, jinyounghan}@skku.edu

Abstract

Discovering individuals' suicidality on social media has become increasingly important. Many researchers have studied to detect suicidality by using a suicide dictionary. However, while prior work focused on matching a word in a post with a suicide dictionary without considering contexts, little attention has been paid to how the word can be associated with the suicide-related context. To address this problem, we propose a suicidality detection model based on a graph neural network to grasp the dynamic semantic information of the suicide vocabulary by learning the relations between a given post and words. The extensive evaluation demonstrates that the proposed model achieves higher performance than the state-of-the-art methods. We believe the proposed model has great utility in identifying the suicidality of individuals and hence preventing individuals from potential suicide risks at an early stage.

1 Introduction

Suicide has become a serious problem in society. The OECD (Organization for Economic Cooperation and Development) reported that the suicide rate of South Korea and the USA was 23.0 and 14.5 deaths per 100,000 population in 2017, which ranked 1st and 8th, respectively[1].

The awareness of the severity of suicide has led researchers to develop suicidality detection models using a deluge of user activity data on social media, which can help capture latent warning signs of suicide in an early stage (Sawhney et al., 2020; Lee et al., 2020; Shing et al., 2020). For example, the prior work showed that linguistic characteristics revealed in social media posts could be linked to suicide risks (De Choudhury et al., 2016; Shing et al., 2018). Specifically, applying the lexicon-based methods using suicide dictionaries made by domain experts has been reported as effective in capturing linguistic characteristics to detect suicidality. (Gaur et al., 2019; Lv et al., 2015).

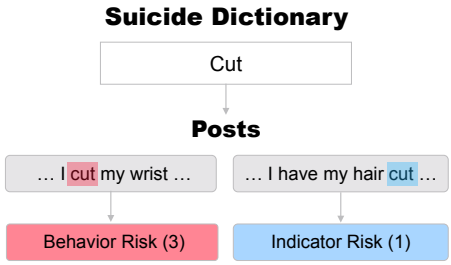

Figure 1: An example of how a word in a suicide dictionary can be misleading in prior work.

While applying the lexicon-based method has been known to be explainable and easy to implement (Kotelnikova et al., 2021; Razova et al., 2021), it may have a limitation: only focusing on whether each word in a post is matched with the suicide lexicon, not considering the context. For example, as illustrated in Figure 1, there are two sentences: "I cut my wrist" and "I have my hair cut". Assuming that the word 'cut' belongs to the suicide dictionary, only the former sentence should be evaluated as having suicidality. However, the latter sentence could also appear to have suicidality if the methods of prior work (Lv et al., 2015; Gaur et al., 2019) are applied. In other words, if the context is incorrectly captured, a model using a suicide lexicon created by experts may not be able to accurately assess the risk of suicidality (Limsopatham and Collier, 2016).

To address this problem, we propose to model the dynamic semantic knowledge between posts and multiple suicide-related words in a suicide dictionary. Capturing the posts' document-word association and word co-occurrence is crucial to un-

*Corresponding author.
[1] https://data.oecd.org/healthstat/suicide-rates.htm

Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, pages 116 - 125
July 15, 2022 ©2022 Association for Computational Linguistics

derstanding the contextualized suicidality revealed in social media posts. To this end, we apply a graph neural network to jointly learn word and document embeddings over a contextual graph representing the relations between posts and multiple suicide-related words in the dictionary. We build a heterogeneous network describing the relations (i) between social media posts and multiple words in a suicide dictionary and (ii) between suicide words based on the co-occurrence. As node information in the given graph, a post node includes the contextual representation obtained from pre-trained BERT (Devlin et al., 2018), and a word node contains the suicide risk level information and contextual representation obtained from the fine-tuned Word2Vec (Mikolov et al., 2013). We learn the proposed heterogeneous graph using the modified GraphSAGE (Hamilton et al., 2017), *Contextual GraphSAGE (C-GraphSAGE)*, to derive a contextualized graph representation.

Instead of using existing suicide dictionaries, we create a word-level suicide dictionary based on social media data using a computational method (Section 3). Since the existing suicide-related lexicon mostly consists of clinical terms (e.g., 'Suicide by self-administered drug') validated by domain experts (Gaur et al., 2019), it may result in a discrepancy with the language used in social media. The created suicide dictionary consists of 279 words and four categories of suicidality levels.

We summarize our contributions as follows.

- We propose a contextualized suicidality detection model *Contextual GraphSAGE (C-GraphSAGE)* using a graph neural network, which can effectively utilize a suicide dictionary. Our evaluation of the real-world dataset demonstrates that the proposed model outperforms the state-of-the-art methods for detecting suicide risk levels using a suicide dictionary.

- We make a word-level English suicide dictionary based on social media data publicly available [2]. We believe the created dictionary can be useful for researchers who want to assess suicidal ideation on social media to prevent potential suicide risks at an early stage.

[2] https://sites.google.com/view/daeun-lee/dataset

2 Related Work

2.1 Suicidality Assessment with Suicide Lexicon

Researchers have investigated that user activity data on social media can provide a cue for analyzing individual suicidality (De Choudhury et al., 2016; Shing et al., 2018). Specifically, prior research showed that linguistic characteristics revealed in social media posts (Sawhney et al., 2020, 2021a) could be linked to suicidal ideation. In particular, utilizing suicide dictionaries made by domain experts has been demonstrated as effective (Lv et al., 2015; Cao et al., 2019; Gaur et al., 2019; Lee et al., 2020), and such lexicon-based methods are known to be fast, explainable, and easy to implement (Kotelnikova et al., 2021; Razova et al., 2021). For example, Lv et al. (2015) developed and validated that a Chinese suicide dictionary made by domain experts helps predict suicidality. Similarly, Gaur et al. (2019) demonstrated the predictive power of suicide dictionaries with domain knowledge.

With the recent advancement of deep learning technologies, high-performing deep learning models have been proposed for accurately assessing suicidality (Sawhney et al., 2021a,b; Cao et al., 2020). In this way, incorporating a suicide dictionary into a deep learning model has received great attention (Cao et al., 2019; Lee et al., 2020). For example, Cao et al. (2019) built suicide-oriented word embeddings to intensify the sensibility of suicide-related lexicons and employed a two-layered attention mechanism. Lee et al. (2020) proposed a deep learning method to utilize existing suicide dictionaries for the low-resource language where a knowledge-based suicide dictionary has not yet been developed. However, the prior work focused on how each word in a post is associated with the words/phrases in a suicide dictionary, e.g., via lexical matching (Lv et al., 2015; Gaur et al., 2019) or fixed word embeddings (Cao et al., 2019; Lee et al., 2020), which may fail to capture the semantic information of suicide lexicons in the suicide-related context.

2.2 Suicidality Assessment with Graph Neural Networks

Among the recent deep learning technologies, graph neural networks (GNNs) have received growing attention in the suicidality assessment task. In particular, GNNs were adopted to extract social

information from a user's neighborhood in a social network formed between different users posting about suicidality (Sinha et al., 2019; Sawhney et al., 2021b). Furthermore, Cao et al. (2020) built personal knowledge graphs on Sina Weibo to utilize rich social interaction data in suicidal ideation detection. Since capturing the posts' document-word association and word co-occurrence is crucial to understanding the contextualized suicide intent revealed in social media posts using the suicide dictionary, we apply a GNN to jointly learn word and document embeddings over a textual graph representing the relations between posts and multiple suicide-related words in the dictionary. Note that GNN has been explored to be useful in jointly learning word and document embeddings over a textual graph representation from the perspective of using lexicon for many NLP tasks (Yao et al., 2019; Tang et al., 2020).

3 Suicide Dictionary

A suicide-related word list can help build a simple detector that automatically responds with helpline links to suicidal content. However, the existing English suicide-related lexicon[3] mainly was made of clinical terms validated by domain experts (Gaur et al., 2019), which results in the discrepancy with the language used in social media. Hence, the authors (Gaur et al., 2019) just used the suicide lexicon as a criterion for checking the presence of a concept in the user's posts. Instead of using the existing English suicide lexicon mostly consisting of clinical terms, we propose to create a word-level English suicide dictionary based on social media data. The proposed computational method can be easily applied to other languages that do not have their own suicide lexicons.

Creating a Suicide Dictionary. We create a word-level English suicide dictionary in a computational way using the UMD Reddit Suicidality Dataset (Shing et al., 2018; Zirikly et al., 2019).

The dataset contains 79,569 posts uploaded to 37,083 subreddits of 866 Reddit users posted on the r/SuicideWatch subreddit from 2008 to 2015. In addition, each post is labeled the suicidality severity conducted by crowdsourcing and domain experts (i.e., No risk, Low risk, Moderate risk, and Severe risk). We only use the posts uploaded to the r/SuicideWatch and 15 mental-health-related subreddits (e.g., r/depression, r/anxiety, r/selfharm, etc.) (Gaur et al., 2018) as a target group and use the posts of users who had not posted on either r/SuicideWatch or mental-health related subreddits as a control group.

[3] https://github.com/AmanuelF/Suicide-Risk-Assessment-using-Reddit

Risk Level	# of Words	Examples
No Risk	55	mother, friend, hope, hug, talk
Low Risk	48	emptiness, overthink, stress, desperate
Moderate Risk	83	scared, lonely, psychiatric, pain
Severe Risk	111	cutting, die, hallucination, dread

Table 1: Example words of the generated suicide dictionary.

Before constructing a dictionary, we anonymize the dataset by removing personally identifiable information such as names, email addresses, and URLs. After removing stopwords and lemmatizing the text using spaCy (Honnibal and Montani, 2017), we extract keywords for each post using KeyBERT (Grootendorst, 2020), and then apply the sparse additive generative model (SAGE) (Eisenstein et al., 2011) to determine the words specialized for each label compared to the entire lexicon. Finally, the constructed dictionary includes 297 suicide-related words. Note that the words belonging to the control group are excluded from the corpus set of each label.

Validation and Correction. We recruited two clinical psychotherapists and a psychiatrist to validate and correct the computationally generated suicide dictionary. All annotators verify how well each label of the suicide word complies with the existing sharing task guideline (Shing et al., 2018; Zirikly et al., 2019), and correct it if it does not meet the criteria. Each annotator performs the validation process independently. The final risk label of each suicide word is set to the label agreed by more than or equal to two annotators. As a result of removing 18 differently validated words from all three annotators, there are 279 words in the final dictionary. Table 1 describes the example of words for each class in the generated suicide dictionary.

4 The Model

We propose a suicidality detection model *C-GraphSAGE* that can capture the severity of suicidality of a post on social media. Figure 2 il-

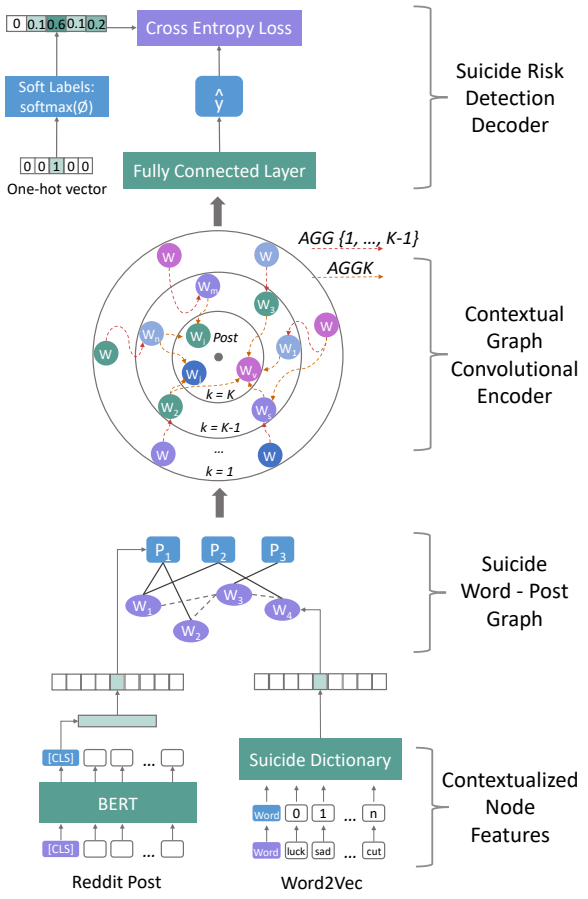

Figure 2: The overall architecture of the model.

lustrates the overall architecture of the proposed model. The model first takes a heterogeneous network that includes posts and suicide words as input. We then apply GraphSAGE (Hamilton et al., 2017) to the given graph to learn the informative representation of suicide-related context by capturing (i) post-words associations and (ii) relations between suicide-related words. Finally, the extracted node presentation from the network is fed into the classification layer. The given post is classified into one of five risk categories: Support (*SU*), Indicator (*IN*), Ideation (*ID*), Behavior (*BR*), and Attempt (*AT*).

4.1 Heterogeneous Network

We build a heterogeneous graph $G = (V_P \cup V_W, E_{PW} \cup E_{WW})$ to represent the relations between social media posts $\{p_i\}_{i=1}^{m} \in P$ and multiple words in a suicide dictionary $\{w_i\}_{i=1}^{n} \in W$, where m and n indicate the number of posts and suicide words, respectively. A graph G consists of two types of nodes, post V_P and suicide word V_W nodes, and two types of edges, post-word E_{PW} and word-word E_{WW} edges. An edge in E_{PW} is linked between a post and its corresponding word if a post contains a specific word in the dictionary. Note that no weight is attached on E_{PW}. An edge in E_{WW} is linked if two words in the suicide dictionary occur together in a post in the UMD Reddit Suicidality Dataset (Shing et al., 2018; Zirikly et al., 2019), which is utilized in constructing a suicide dictionary (in Section 3). A weight on an edge in E_{WW} can be computed by the positive Point-wise Mutual Information (PMI) score that can capture collocations and relations between two terms (Yao et al., 2019; Tang et al., 2020) as follows:

$$PMI(w_i, w_j) = log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \quad (1)$$

Note that we only attach the edge weight on a suicide word pair with the positive PMI value, which indicates a high semantic correlation of two words in a document.

Contextualized Node Features. In order to generate node features of posts X_P and suicide words X_W, we employ the pre-trained BERT for posts and pre-trained Word2Vec for suicide words, respectively, to capture the contextual representation of text features. Specifically, to obtain X_P, a post p is fed into the BERT model and obtain the `[CLS]` token as a sentence-level representation of the claim as follows:

$$X_{p_i} = BERT(p_i) \in \mathbb{R}^{1 \times d_{cls}} \quad (2)$$

where d_{cls} is the dimension size of a contextualized embedding of `[CLS]` and p_i is i^{th} post. For representing each suicide word w_i, we apply the word-embedding from the pre-processed texts using the Word2Vec model, Gensim (Rehurek and Sojka, 2010). The word vectors are pre-trained with the Skip-Gram representation model using the UMD Reddit Suicidality Dataset (Shing et al., 2018; Zirikly et al., 2019), while the size of the window and the dimension are set to 5 and 200, respectively. Finally, X_W is (i) the suicide risk level (i.e., 0, 1, 2, 3) of each word RL_W and (ii) word embeddings WV_W from pre-trained Word2Vec as follows:

$$RL_{w_i} = \begin{cases} 3, & \text{Severe Risk} \\ 2, & \text{Moderate Risk} \\ 1, & \text{Low Risk} \\ 0, & \text{No Risk} \end{cases} \quad (3)$$

$$WV_{w_i} = Word2Vec(w_i) \in \mathbb{R}^{1\times d_{wv}} \quad (4)$$

$$X_{w_i} = RL_{w_i} \oplus WV_{w_i} \in \mathbb{R}^{1\times(d_{wv}+1)} \quad (5)$$

where d_{wv} is the dimension size of a Word2Vec and w_i is i^{th} word in the suicide dictionary.

4.2 Contextualized Graph Convolutional Encoder

To generate node embedding from the given heterogeneous graph model, we apply the GraphSAGE (Hamilton et al., 2017), a well-known model for a graph neural network (GNN) that supports batch-training without updating states over the whole graph and has shown experimental success compared to other graph representation learning models (Tang et al., 2020). The model first recursively updates embedding for each node v from V_P and V_W by aggregating information from node v's immediate neighbors $N(v), u \in N(v)$, through the aggregation function at each search depth k. After that, h_v^k, node v's representation at step k, is updated by combining h_v^{k-1} and the information obtained from $h_{\mathcal{N}(v)}^{(k)}$, which is the representation of v's neighboring nodes at step k. As suggested in Hamilton et al. (2017), the neighboring nodes are uniformly sampled with a fixed-size set for each search depth. The initial output is $h_v^0 = X_v$. The series of updating processes is defined as follows.

$$h_{\mathcal{N}(v)}^{(k)} = \text{aggregate}_\text{k}\left(\{h_u^{k-1}, \forall u \in \mathcal{N}(v)\}\right) \quad (6)$$

$$h_v^{(k)} = \sigma\left(W^k \cdot \text{concat}(h_v^{k-1}, h_{\mathcal{N}(v)}^k)\right)) \quad (7)$$

As shown in Figure 3, we propose to use an aggregation function (Eq. 6) based on a convolutional neural network (CNN) instead of existing aggregators such as pool, LSTM, and mean, used in Hamilton et al. (2017). A CNN is proven to be effective in detecting local patterns (Minaee et al., 2021), hence it generates a feature map over the neighbor node embeddings that can explicitly capture relations of words in the suicide dictionary.

Given the target node v's neighboring nodes $\{u_i\}_{i=1}^j \in \mathcal{N}(v)$' embedding $\left\{h_{u_1}^{k-1}, h_{u_2}^{k-1}, \cdots, h_{u_j}^{k-1}\right\} \in \mathbb{R}^{j\times d}$, where d is the dimension of node feature, a convolution operation involving a filter $q \in \mathbb{R}^{l*d}$ generates a feature c_i from a window of nodes $u_{i:i+l-1}$ as follows.

$$c_i = \sigma(q \cdot u_{i:i+l-1} + b) \quad (8)$$

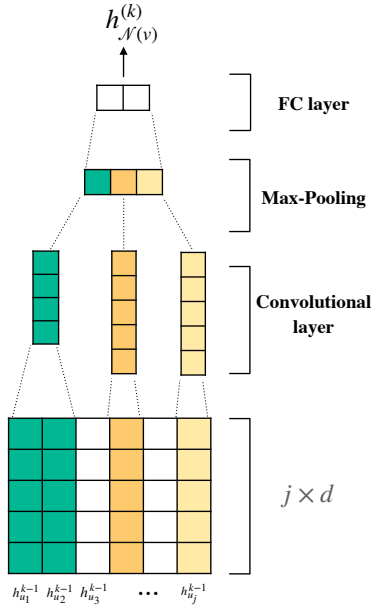

Figure 3: The example of aggregating information from neighborhood of the target node by CNN.

where b is a bias term and $ReLu$ (Nair and Hinton, 2010) is adopted as the non-linear function σ. The filter is employed to each possible window of neighboring nodes to produce a feature map as follows.

$$c = [c_1, c_2, \cdots, c_{j-l+1}] \in \mathbb{R}^{j-l+1} \quad (9)$$

To capture the diverse local structure, we adopt multiple filters with different sizes. For example, the set of kernel sizes used in this paper is [1,2,3]. In this way, the filter can create up to 3 neighbor nodes' combinations. We then apply a max-pooling operation (Collobert et al., 2011) over the feature map and take the maximum value $\hat{c} = max\{c\}$ as the feature corresponding to the filter. Finally, we derive a node v's neighbor nodes' representation as follows.

$$h_{\mathcal{N}(v)}^{(k)} = \mathcal{F}_c(\hat{c}) \in \mathbb{R}^{1\times d} \quad (10)$$

Note that, if node v has neighbors with different node types, we sum representations of neighbor nodes. Since we predict the suicidality level of the post, we only consider the node V_p's representation.

4.3 Suicidality Detection Decoder

To predict the suicidality level of a post, the proposed decoder identifies suicidal severity for each node by learning the graph representation as follows.

$$\hat{y} = \mathcal{F}_c(h_v^{(k)}) \quad (11)$$

Like Sawhney et al. (2021a), we adopt the ordinal regression loss (Diaz and Marathe, 2019) as an objective function. Instead of using an one-hot vector representation of the true labels, they used a soft encoded vector representation by considering the ordinal nature between suicidality levels. While ground truth labels are denoted as $\mathcal{Y} = \{SU = 0, IN = 1, ID = 2, BR = 3, AT = 4\} = \{r_i{}_{i=0}^{4}\}$, soft labels as probability distributions of ground truth labels is denoted by y = $[y_0, y_1, y_2, y_3, y_4]$. The probability y_i of each risk-level r_i is

$$y_i = \frac{e^{-\phi(r_t, r_i)}}{\sum_{k=1}^{\lambda} e^{-\phi(r_t, r_i)}} \forall r_i \in \mathcal{Y} \tag{12}$$

where $e^{-\phi(r_t, r_i)}$ is a cost function that penalizes how far the true risk-level r_t is from a risk-level $r_i \in \mathcal{Y}$, which is formulated as $e^{-\phi(r_t, r_i)} = \alpha |r_t - r_i|$, where α is a penalty parameter for incorrect prediction.

Finally, the cross-entropy loss is calculated using the probability distribution y and classification score $\hat{y}$ obtained in Eq(11) as follows:

$$\mathcal{L} = -\frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{\lambda} y_{ij} \log \hat{y}_{ij} \tag{13}$$

where n is the batch size and λ is the number of risk-levels.

5 Experiments

We evaluate the our proposed model by answering the following research questions:

- RQ1: Is the proposed suicide dictionary made by a computational method effective in detecting suicidality risk?
- RQ2: Can using the suicide dictionary help improve the model performance?
- RQ3: Is the C-GraphSAGE efficient in utilizing the suicide dictionary?

5.1 Dataset

To learn our proposed model, we utilize *The Golden Standard Dataset* introduced by (Gaur et al., 2019), which consists of Reddit posts collected from the 9 suicide-related subreddits (e.g., r/SuicideWatch and r/depression). The dataset is within the time frame from 2005 to 2016 and annotated with 5 suicidality levels (i.e., Supportive, Indicator, Ideation, Behavior, and Attempt) by mental health experts[4]. While the dataset contains both user-level and post-level data, we utilize the post-level data in this paper since our model aims to detect suicidality levels for a given social media post, and a post-level prediction can be useful for immediate or early intervention on suicidality risks. Finally, the dataset includes 1346, 420, 337, 77, and 49 posts for the Supportive, Indicator, Ideation, Behavior, and Attempt levels, respectively. In addition, we implement a stratified 60:20:20 split such that the train, validation, and test sets consist of 1,427, 356, and 446 posts, respectively.

5.2 Evaluation Metrics

To consider the ordinal nature of suicidality risk levels, we adopt the modified definitions of False Positive (FP), False Negative (FN) (Gaur et al., 2019) in our experiments as follows.

$$FP = \frac{\sum_{i=1}^{N_T} I(\hat{y}_i > y_i)}{N_T} \tag{14}$$

$$FN = \frac{\sum_{i=1}^{N_T} I(y_i > \hat{y}_i)}{N_T} \tag{15}$$

where $\hat{y}_i$ is the predicted level, y_i is the actual level for i^{th} test data, and N_T is the size of the test data. $\Delta(y_i, \hat{y}_i)$ is the difference between y_i and $\hat{y}_i$. The evaluation metric terms for precision and recall are renamed as graded precision and graded recall, respectively.

5.3 Baselines and Experiment Settings

We compare the proposed model against the following three types of models: (1) Lexicon-based approaches; Rule-based (Gaur et al., 2019), SVM (Lv et al., 2015), and Random Forest (RF) (Amini et al., 2016), (2) Deep learning approaches w/o lexicon; Contextual CNN (Gaur et al., 2019), SISMO (Sawhney et al., 2021a), and BERT (Devlin et al., 2018), and (3) Lexicon + deep learning; Cao et al. (2019) and Reformed BERT. Detailed experimental settings for reproducibility are summarized in the Appendix **??**.

We tune hyperparameters based on the highest FScore obtained from the validation set for all the models. We use the grid search to explore (i) the number of kernel output size in aggregate function $\tilde{q}$, (ii) the number of post features in hidden state H^D, (iii) the initial learning rate lr, and (iv) the dropout rate σ. The optimal hyperparameters were

[4] https://github.com/AmanuelF/Suicide-Risk-Assessment-using-Reddit

Type of Model	Model	Loss	G-Precision	G-Recall	G-F1
Suicide lexicon only	Rule-based (Gaur et al., 2019)	/	0.33	0.74	0.46
	SVM (Lv et al., 2015)	Hinge Loss	0.51	0.66	0.58
	RF (Amini et al., 2016)	Gini Impurity	0.65	0.67	0.66
Deep learning only	Contextual CNN (Gaur et al., 2019)	Cross Entropy	0.78	0.57	0.66
	SISMO (Sawhney et al., 2021a)	Soft Label	0.77	0.77	0.77
	SDM w/o Lexicon (Cao et al., 2019)	Cross Entropy	0.73	0.75	0.74
	BERT w/o Lexicon (Devlin et al., 2018)	Soft Label	0.81	0.80	0.80
Suicide lexicon + Deep learning	SDM w/ Lexicon (Cao et al., 2019)	Cross Entropy	0.75	0.78	0.77
	BERT w/ Lexicon (Devlin et al., 2018)	Soft Label	0.82	0.79	0.81
	C-GraphSAGE (Ours)	Soft Label	**0.85**	**0.82**	**0.84**

Table 2: Performance comparisons of the proposed model and baselines.

found to be: $\tilde{q} = 50, \tilde{H^D} = 512, lr = 3e-5$, and $\sigma = 0.1$.

6 Results

In this section, we present our experiment results to answer the three above research questions. Table 2 summarizes the overall performance results of the proposed model (C-GraphSAGE) and the baselines.

6.1 RQ1: Is the proposed suicide dictionary made by a computational method effective in detecting suicidality risks?

Model	Lexicon	Precision	Recall	FScore
Rule-Based (Gaur et al., 2019)	Gaur et al. (2019)	0.26	0.70	0.38
	Ours	**0.33**	**0.74**	**0.46**
RF	Gaur et al. (2019)	0.51	0.65	0.57
	Ours	**0.65**	**0.67**	**0.66**

Table 3: Performance Comparisons between the existing suicide dictionary made by domain experts and the proposed computationally created dictionary (Ours).

To answer the first question, we evaluate the suicidality detection models (Rule-based (Gaur et al., 2019) and Random Forest (RF)) with two different suicide dictionaries: (1) the domain knowledge-based one made by experts (Gaur et al., 2019), and (2) a computationally created one (Ours). As shown in Table 3, the performance with the suicide dictionary created by a computation method (Ours) outperforms the domain knowledge-based lexicon. Furthermore, it indicates that a word-level English suicide dictionary based on social media data is helpful to be mapped with social media posts for detecting suicidality. In other words, the proposed computational method to create a suicide dictionary effectively detects suicidality.

6.2 RQ2: Can using the suicide dictionary help improve the model performance?

Overall, deep learning models with a suicide dictionary (i.e., C-GraphSAGE, 'SDM w/ lexicon', and 'BERT w/ lexicon') perform better than the models that use only text information such as C-CNN, SISMO, 'SDM w/o lexicon', and 'BERT w/o lexicon'. This shows that a model using a suicide dictionary can present the suicide-related context of posts, resulting in high performance. Note that 'SDM w/ lexicon' uses the fine-tuned word embedding model to capture domain knowledge from a pre-built suicide dictionary (Cao et al., 2019), whereas 'SDM w/o lexicon' adopts pre-trained FastText embeddings (Bojanowski et al., 2017) for encoding posts. Also, 'the BERT w/ lexicon' adds the suicide words on the BERT-Tokenizer.

6.3 RQ3: Is the C-GraphSAGE efficient in utilizing the suicide dictionary?

C-GraphSAGE outperforms the other model using a suicide dictionary, the Reformed BERT, offering an insight that capturing dynamic semantic information from a suicide dictionary is beneficial rather than considering only the presence of suicide words. We attribute this to the strength of the graph neural network model that can learn better representations from the relations between posts and words in the suicide dictionary and the associations between suicide words in the suicide-related context. As a result, C-GraphSAGE is helpful in accurately identifying suicidality levels, which shows outstanding utility in preventing suicide risks.

6.4 Ablation Study

We perform an ablation study to examine the effectiveness of different aggregation functions over the proposed C-GraphSAGE, as shown in Table 4. We compare the proposed CNN-based aggregation

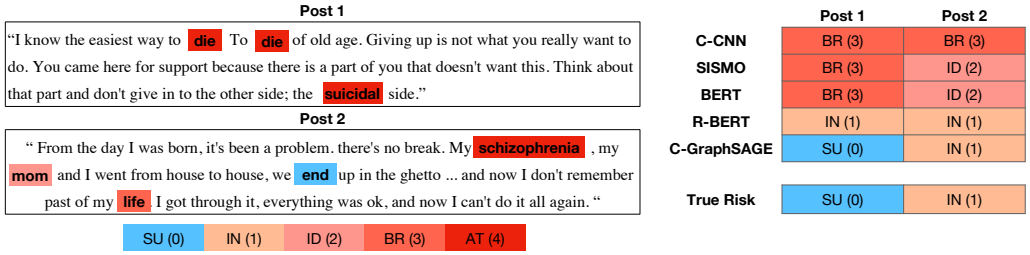

Figure 4: A qualitative analysis on the two cases shows the C-GraphSAGE can capture the risk levels accurately.

Aggregation Function	G-Precision	G-Recall	G-F1
C-GraphSAGE + Pool	0.81	0.79	0.80
+ LSTM	0.81	0.79	0.80
+ MEAN	0.87	0.78	0.82
+ biLSTM	0.88	0.78	0.83
+ CNN (Ours)	0.85	**0.82**	**0.84**

Table 4: An ablation study on different aggregation functions over C-GraphSAGE.

function with the three popular aggregation functions $\in \{LSTM, Pool, Mean\}$ (Hamilton et al., 2017) as well as $bi - LSTM$ (Tang et al., 2020). As shown in Table 4, the model performance significantly improves when we use the aggregation function based on a CNN than other aggregators. Notably, the CNN aggregator outperforms the biLSTM (Tang et al., 2020). This is because an RNN works well in capturing long-term dependencies, whereas a CNN can effectively identify structural patterns. In other words, it is crucial to capture local relations between words than the order of words in our case. We believe that the proposed aggregator can effectively capture neighboring node information, thereby enhancing the robustness of the model for unseen data.

6.5 Qualitative Analysis

To provide detailed insight and interpretability, we qualitatively analyze two cases where C-GraphSAGE performs better than other models in Figure 4. We compare how to predict suicidality by each model given the input that contains the same suicide words. Both posts contain high-level suicide words, but the actual suicidality is relatively low. The proposed model C-GraphSAGE predicts the corresponding risk accurately, whereas other models that assess risk only by the presence of suicide words are likely to classify suicidality levels more highly than actual levels.

7 Concluding Discussion

This paper proposed a suicidality detection model, C-GraphSAGE, which can capture the context of suicidality by learning the relations between social media posts and suicide-related words. Using a word-level English suicide dictionary validated by domain experts, the proposed model achieved higher performance than the state-of-the-art methods in detecting suicidality levels. We believe the proposed model has great utility in identifying potential suicidality levels of individuals with social media data, preventing individuals from potential suicide risks at an early stage.

Ethical Concerns. This study is reviewed and approved by the Institutional Review Board (SKKU2020-10-021). All datasets are anonymized. Hence no personal information can be identifiable.

Limitation. Assessing suicidality using social media data is subjective (Keilp et al., 2012), and the analysis of this paper can be interpreted in diverse ways across the researchers. The experiment data may be sensitive to demographic, annotator, and media-specific biases (Hovy and Spruit, 2016). The analytical patterns learned by C-GraphSAGE may fail to generalize to other social media due to the relatively small data and/or short time window appeared in Reddit. Nevertheless, an interpretable model can help to follow and improve other targets with different statistical patterns and biases (Jacobson et al., 2020).

There is an overlap in data collection periods between the data used to create the suicide dictionary (2008 – 2015) and the data used in the experiment (2005 – 2016). Since all the datasets are anonymized, a Jaccard similarity analysis (Jaccard, 1908) is performed in a grid manner to determine a similarity between all post pairs in two datasets. The result shows that the Jaccard coefficient is quite low (max = 0.5 , mean = 0.1, std = 0.05), meaning that both groups are unrelated.

Practical Applicability. The proposed suicidality detection model can be used for screening or identifying individuals at risk on social media to prioritize early intervention for clinical support.

Acknowledgments

This research was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2022S1A5A8054322), and the MSIT (Ministry of Science and ICT), Korea, under the ICAN (ICT Challenge and Advanced Network of HRD) program (IITP-2021-2020-0-01816) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation).

References

Payam Amini, Hasan Ahmadinia, Jalal Poorolajal, and Mohammad Moqaddasi Amiri. 2016. Evaluating the high risk groups for suicide: a comparison of logistic regression, support vector machine, decision tree and artificial neural network. *Iranian journal of public health*, 45(9):1179.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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

Lei Cao, Huijun Zhang, Ling Feng, Zihan Wei, Xin Wang, Ningyun Li, and Xiaohao He. 2019. Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1718–1728.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Raul Diaz and Amit Marathe. 2019. Soft labels for ordinal regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4738–4747.

Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1041–1048. Citeseer.

Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware assessment of severity of suicide risk for early intervention. In *proceedings of the 2019 World Wide Web Conference*, pages 514–525.

Manas Gaur, Ugur Kursuncu, Amanuel Alambo, Amit Sheth, Raminta Daniulaityte, Krishnaprasad Thirunarayan, and Jyotishman Pathak. 2018. " let me tell you about your mental health!" contextualized classification of reddit posts to dsm-5 for web-based intervention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 753–762.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

William L Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1025–1035.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Paul Jaccard. 1908. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, 44:223–270.

Nicholas C Jacobson, Kate H Bentley, Ashley Walton, Shirley B Wang, Rebecca G Fortgang, Alexander J Millner, Garth Coombs III, Alexandra M Rodman, and Daniel DL Coppersmith. 2020. Ethical dilemmas posed by mobile health and machine learning in psychiatry research. *Bulletin of the World Health Organization*, 98(4):270.

John G Keilp, Michael F Grunebaum, Marianne Gorlyn, Simone LeBlanc, Ainsley K Burke, Hanga Galfalvy, Maria A Oquendo, and J John Mann. 2012.

Suicidal ideation and the subjective aspects of depression. *Journal of affective disorders*, 140(1):75–81.

Anastasia Kotelnikova, Danil Paschenko, Klavdiya Bochenina, and Evgeny Kotelnikov. 2021. Lexicon-based methods vs. bert for text sentiment analysis. *arXiv preprint arXiv:2111.10097*.

Daeun Lee, Soyoung Park, Jiwon Kang, Daejin Choi, and Jinyoung Han. 2020. Cross-lingual suicidal-oriented word embedding toward suicide prevention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2208–2217.

Nut Limsopatham and Nigel Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1014–1023.

Meizhen Lv, Ang Li, Tianli Liu, and Tingshao Zhu. 2015. Creating a chinese suicide dictionary for identifying suicide risk on social media. *PeerJ*, 3:e1455.

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

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: A comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Icml*.

Elena Razova, Sergey Vychegzhanin, and Evgeny Kotelnikov. 2021. Does bert look at sentiment lexicon? *arXiv preprint arXiv:2111.10100*.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*. Citeseer.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Shah. 2020. A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7685–7697.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2021a. Towards ordinal suicide ideation detection on social media. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 22–30.

Ramit Sawhney, Harshit Joshi, Rajiv Shah, and Lucie Flek. 2021b. Suicide ideation detection via social and temporal user representations using hyperbolic learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2176–2190.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36.

Han-Chin Shing, Philip Resnik, and Douglas W Oard. 2020. A prioritization model for suicidality risk assessment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8124–8137.

Pradyumna Prakhar Sinha, Rohan Mishra, Ramit Sawhney, Debanjan Mahata, Rajiv Ratn Shah, and Huan Liu. 2019. # suicidal-a multipronged approach to identify and explore suicidal ideation in twitter. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 941–950.

Pingjie Tang, Meng Jiang, Bryan Ning Xia, Jed W Pitera, Jeffrey Welser, and Nitesh V Chawla. 2020. Multi-label patent categorization with non-local attention-based graph convolutional network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9024–9031.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7370–7377.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 24–33.