Cross-Lingual Suicidal-Oriented Word Embedding toward Suicide Prevention

Daeun Lee[1], **Soyoung Park**[2], **Jiwon Kang**[1], **Daejin Choi**[3], **Jinyoung Han**[1*]
[1]Department of Applied Artificial Intelligence, Sungkyunkwan University
[2]National Assembly Research Service
[3]Department of Computer Science & Engineering, Incheon National University
{delee12, jiwonkang, jinyounghan}@skku.edu
sypark@assembly.go.kr, djchoi@inu.ac.kr

Abstract

Early intervention for suicide risks with social media data has increasingly received great attention. Using a suicide dictionary created by mental health experts is one of the effective ways to detect suicidal ideation. However, little attention has been paid to validate whether and how the existing dictionaries for other languages (i.e., English and Chinese) can be used for predicting suicidal ideation for a low-resource language (i.e., Korean) where a knowledge-based suicide dictionary has not yet been developed. To this end, we propose a cross-lingual suicidal ideation detection model that can identify whether a given social media post includes suicidal ideation or not. To utilize the existing suicide dictionaries developed for other languages (i.e., English and Chinese) in word embedding, our model translates a post written in the target language (i.e., Korean) into English and Chinese, and then uses the separate suicidal-oriented word embeddings developed for English and Chinese, respectively. By applying an ensemble approach for different languages, the model achieves high accuracy, over 87%. We believe our model is useful in accessing suicidal ideation using social media data for preventing potential suicide risk in an early stage.

1 Introduction

As online social media has become the norm to share our daily lives, people often share their emotions, feelings, and mental state. This has spurred scholars to identify diverse mental health problems such as depression, anxiety, bipolar disorder, or suicidal thoughts using plenty of user behavior data on online social media (Ji et al., 2019; Pavalanathan and De Choudhury, 2015; Kim et al., 2020). Such user behavior data can provide a cue for identifying individual mental state or even suicide risk (O'dea et al., 2015; Ren et al., 2015; Coppersmith et al., 2018), which can be used to support mental health care (Shen and Rudzicz, 2017; Suhara et al., 2017).

Among the diverse mental health problems, suicide has become one of the big and emerging concerns worldwide. The OECD (Organization for Economic Cooperation and Development) reported 11.2 deaths per 100,000 population in OECD countries in 2017 (OECD, 2020). In particular, the suicide rate of Korea and the USA was 24.6 and 13.9 deaths per 100,000 population in 2016, which ranked 1st and 8th, respectively.

The awareness of the severity of suicide has led researchers to assess mental health using social media data for recognizing potential warning signs of suicide in an early stage (Pavalanathan and De Choudhury, 2015; O'dea et al., 2015). In particular, linguistic characteristics (e.g., frequently used words like 'family', 'sad', or 'dream') of social media posts have been extensively investigated (Gaur et al., 2019; Lv et al., 2015). As prior research showed that certain linguistic features revealed in an individual language could be linked to suicide risk (McCarthy, 2010; Sueki, 2015), there have been attempts to develop machine-learning models using a suicide dictionary, which was created and curated by mental health experts. For example, an English suicide dictionary was created and validated by four clinical psychiatrists (Gaur et al., 2019); a Chinese suicide dictionary was curated by eleven mental health experts (Lv et al., 2015).

The predictive power of such suicide dictionaries with domain knowledge (in English or Chinese) in identifying suicide risk from an English- or Chinese-written social media post has been demonstrated (Gaur et al., 2019; Lv et al., 2015). However, little attention has been paid to validate whether the existing dictionary developed for the specific language (e.g., English or Chinese) can be used for predicting suicidal ideation with other languages

[*] Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2208–2217
November 16 - 20, 2020. ©2020 Association for Computational Linguistics

(e.g., Korean or Japanese), where any suicide dictionary has not yet been developed. It is essential to investigate whether and how existing suicide dictionaries developed by domain experts can be utilized by predicting suicidal ideation for non-English or non-Chinese spoken countries because building and validating such a knowledge-based dictionary requires much effort.

To shed light on this issue, we propose a cross-lingual suicidal ideation detection model that can identify whether a given social media post includes suicidal ideation or not. To utilize the existing suicide dictionaries developed for other languages (i.e., English and Chinese) in word embedding, our model translates a post written in the target language (i.e., Korean) into English and Chinese and then uses the separate word embeddings developed for English and Chinese, respectively. Our model then uses attention to make a representation for post embedding. The attention helps find words that are more relevant to suicidal ideation, thereby obtaining a better post representation. By applying an ensemble approach for different languages, which can reflect linguistic or cultural differences (Lin et al., 2018), our proposed model finally predicts suicidal ideation of the given post in Korean.

We highlight the main contributions of our work as follows.

- To the best of our knowledge, this is the first attempt to utilize the suicide dictionaries developed for other languages (i.e., English and Chinese) in predicting suicidal ideation in Korean. We believe the proposed model provides a cost-effective way to detect suicide risk from a social media post written in a low-resource language where a knowledge-based suicide dictionary does not exist. The proposed model achieves high accuracy, over 87%.

- We make the suicidal-oriented word-embeddings in Korean, English, and Chinese publicly available at `https://dsail-skku.github.io/Cross-Lingual-Suicidal-Embedding/`. Note that the Korean suicidal-oriented word-embedding is built by a computational approach without medical knowledge base but shows a considerable performance in suicidal ideation detection. We believe the suicidal-oriented word-embeddings can be useful for researchers who want to access suicidal ideation using social media data for preventing potential suicide risk at an early stage.

2 Related Work

2.1 Suicide Risk Assessment with Social Media Data

It becomes the norm for people to share their daily lives or feelings on diverse social media. This in turn has led researchers to investigate individuals' mental health problems using a deluge of user activity data on social media (Ji et al., 2019; Pavalanathan and De Choudhury, 2015; Shing et al., 2018), because such user behavior can provide a cue for identifying individual mental state or even suicide risk (O'dea et al., 2015; Ren et al., 2015; Coppersmith et al., 2018; Sinha et al., 2019). There has been great interest in developing a model to detect suicide risks based on user behavior such as the number of posts or followers (Kumar et al., 2015; Cao et al., 2019) and linguistic characteristics (e.g., frequently used words like 'family', 'sad', or 'dream') revealed in social media posts (Gaur et al., 2019; Lv et al., 2015). For example, Coppersmith et al. (2015) conducted a linguistic analysis on social media data and found a few signals that can be linked to suicide attempts and suicidal ideation. De Choudhury et al. (2016) analyzed user post data in Reddit and found that individuals who could become suicidal tend to exhibit changes in linguistic structures, interpersonal awareness, and social interactions in social media. Such identified distinctive markers of shifts can be used for identifying individual suicidal ideation.

2.2 Suicide Dictionary Development

As it has been reported that certain linguistic features revealed in individual language can link suicide risk (McCarthy, 2010; Sueki, 2015), there have been attempts to develop a learning-based model using a suicide dictionary which is created and curated by mental health experts. For example, Gaur et al. (2019) identified and curated English words indicating the severity of suicide risk, resulting in an English suicide dictionary, which was validated by mental health experts. Lv et al. (2015) created and validated a Chinese suicide dictionary with 11 experts, which can be used in predicting individuals' likelihood of suicide. The predictive power of such suicide dictionaries with domain knowledge has been demonstrated (Gaur et al., 2019; Lv et al.,

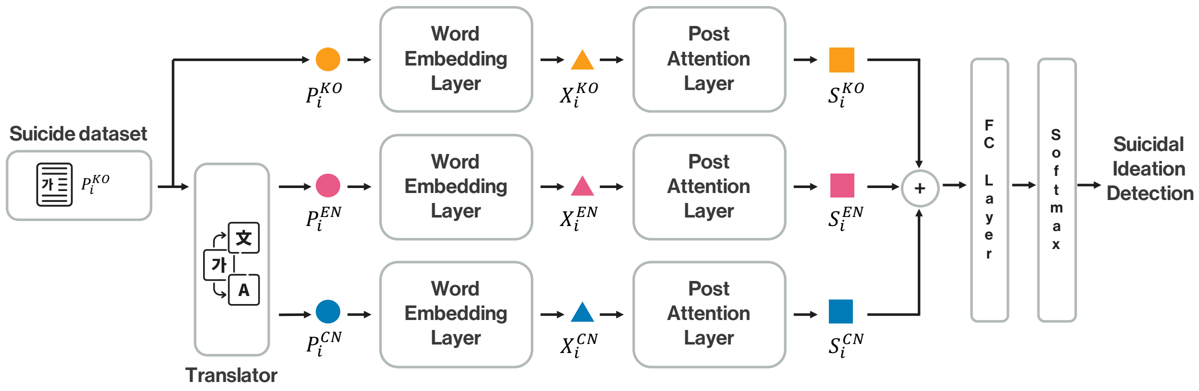

Figure 1: The overall architecture of the cross-lingual suicidal ideation model.

2015). However, little attention has been paid to whether existing dictionaries developed for specific countries or languages (e.g., English or Chinese) can be used for predicting suicidal ideation with other languages such as Korean or Japanese, where any suicide dictionary has not yet been developed. This paper proposes and evaluates a model for predicting suicidal ideation using Korean social media data by exploiting multiple suicide dictionaries developed for other languages (e.g., English and Chinese).

3 Cross-lingual Suicidal Ideation Detection Model

We propose a suicidal ideation detection model that can identify whether a given post includes suicidal ideation or not. To utilize the existing suicide-related dictionaries developed for other languages (i.e., English and Chinese) in word embedding, our model translates a post written in the target language (i.e., Korean) into English and Chinese and then uses the separate word embeddings developed for English and Chinese, respectively. Note that we use Naver Papago (Lee et al., 2016) for translation, which is known to be an efficient translator from Korean to other languages. By applying an ensemble approach for different languages, our proposed model finally predicts suicidal ideation of the given post in Korean. Figure 1 illustrates the overall architecture of our proposed model.

3.1 Suicidal-oriented Word Embedding

We adopt a suicidal-oriented word embedding similar to the prior work (Cao et al., 2019) that refines a word embedding to capture domain knowledge from a pre-built suicide-related dictionary. Figure 2 illustrates the model that identifies whether a given sentence contains suicidal expression or not.

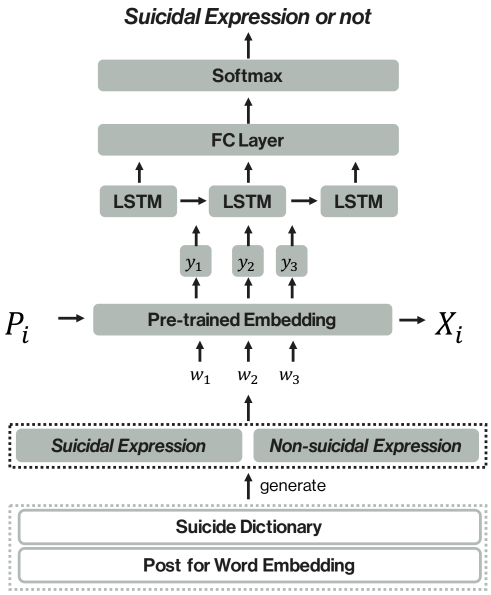

Figure 2: The architecture of the word embedding layer.

3.1.1 Generating suicidal and non-suicidal expressions

For training a suicidal-oriented word embedding, we use a pre-built suicide-related dictionary. If such a dictionary contains word-level information that exhibits how much a word is associated with suicidal ideation (like a Chinese dictionary (Lv et al., 2015)), we apply the word-masking classification method similar to the prior work (Cao et al., 2019). To this end, we generate suicidal and non-suicidal expressions for a given input suicide-related post collected for word embedding, e.g., Weibo Tree Hole data (Cao et al., 2019). The suicidal expression is generated based on the input data itself. For generating a non-suicidal expression, we replace all the suicide-related words (in the dictionary) with "[mask]" in the given input. To avoid learning from the "[mask]" words themselves replaced in the non-suicidal expression, we randomly add two "[mask]" words in the suicidal expression. During the training, we randomly select 50% of the generated suicidal and non-suicidal expressions, respectively, for each epoch.

If a pre-built suicide-related dictionary contains sentence-level information such as *Gold Standard Dataset* (Gaur et al., 2019), which includes English sentences related to suicidal ideation, directly applying the word-masking method (Cao et al., 2019) is not possible; words for masking cannot be extracted from the given sentence-level dictionary. Hence, we use the sentences belonging to the dictionary as suicidal expressions and non-suicide-related posts as non-suicidal expressions for developing the sentence-level word embedding. To generate non-suicidal expressions, we randomly select the (non-suicidal-related) posts on Reddit. Note that the ratio of the generated suicidal and non-suicidal expressions is 1:1.

3.1.2 Word embedding

Given a set of words $A_j = \{w_1, w_2, ..., w_n\}$ in a given expression j labeled as suicidal or non-suicidal, we define $Y_j = \{y_1, y_2, ..., y_n\} \in \mathbb{R}^{n \times d_e}$ is the word embedding of A_j, where n is the length of the words in the given expression j and d_e is the whole dimension of the embedding. Then, each word in Y_j is fed into an LSTM cell to derive text representation:

$$h_t = LSTM(y_t, h_{t-1}) \tag{1}$$

where h_{t-1} and h_t represent the hidden state at time $t-1$ and t, respectively. Note that $H^A = \{h_1, h_2, ..., h_n\} \in \mathbb{R}^{n \times d_e}$ represents a textual representation of A_j. Finally, the model classifies whether the expression is suicidal or not as follows:

$$Softmax((H^A W_1 + b_1)^T W_2 + b_2) \tag{2}$$

where $W_1 \in \mathbb{R}^{d_e \times 1}$, $b_1 \in \mathbb{R}^{1 \times 1}$, $W_2 \in \mathbb{R}^{n \times 2}$ and $b_2 \in \mathbb{R}^{1 \times 2}$ are trainable parameters.

3.2 Post Attention Layer

Given a post p_i, by passing through the corresponding suicidal-oriented word embedding, we obtain the word embedding $X_i = \{x_1, x_2, ..., x_n\} \in \mathbb{R}^{n \times d_e}$ for post p_i. After that, we feed X_i into the LSTM layer as follows:

$$h_t = LSTM(x_t, h_{t_1}) \tag{3}$$

Note that $H^p = \{h_1, h_2, ..., h_n\} \in \mathbb{R}^{n \times d_e}$ is a textual representation of p_i after the LSTM layer.

We then apply the attention mechanism to reflect the important suicide-related information of H^p as follows:

$$\begin{aligned} Attention &= (H^p)^T \times ((H^p)W_3 + b_3) \\ S &= tanh((Attention^T \oplus h_n)W_4 + b_4) \end{aligned} \tag{4}$$

where $Attention \in \mathbb{R}^{1 \times n}$ is the score vector of attention, $S \in \mathbb{R}^{1 \times 32}$ is the final contextual vector, h_n is the last hidden state of the last LSTM cell, and tanh is activation function. $W_3 \in \mathbb{R}^{256 \times 1}$, $b_3 \in \mathbb{R}^{1 \times 1}$,$W_4 \in \mathbb{R}^{512 \times 32}$ and $b_4 \in \mathbb{R}^{1 \times 32}$ are trainable parameters. Figure 3 represents the architecture of the post attention layer.

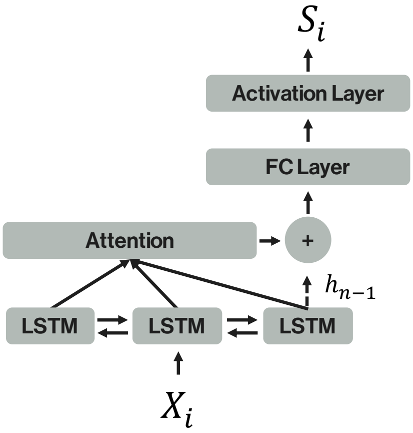

Figure 3: The architecture of the post attention layer.

3.3 Ensemble Layer

For a given set of contextual vectors for different languages, S^{KR}, S^{CN} and S^{EN} in our case, we concatenate them to obtain the total post representation $Q \in \mathbb{R}^{1 \times 96}$:

$$Q = S_i^{KR} \oplus S_i^{CN} \oplus S_i^{EN} \tag{5}$$

A fully-connected layer with an activation function $Relu$ is first applied into Q. We then finally classify whether the post p_i includes suicidal ideation or not as follows:

$$Softmax(Relu(QW_5 + b_5)W_6 + b_6) \tag{6}$$

where $W_5 \in \mathbb{R}^{96 \times 32}$, $b_5 \in \mathbb{R}^{1 \times 32}$, $W_6 \in \mathbb{R}^{32 \times 2}$ and $b_6 \in \mathbb{R}^{1 \times 2}$ are trainable parameters.

Suicide-related post	**Non-suicide-related post**
It's not easy to die.	*So annoyed by the company person.*
I feel like something is pressing me from above. I want to leave.	*Where do you want to go right after Corona?*
I've been on depression medication for years.	*On wet nights, I like the calm song before going to bed.*
I hope my eyes don't open when I lie down to sleep.	*I picked up the money. I feel like something good will come.*

Table 1: Examples of the translated target posts collected from Naver Cafe. Note that the source language is written in Korean.

Language	Chinese	English	Korean
Source	*Chinese Suicide Dictionary* (Lv et al., 2015)	*Gold Standard Dataset* (Gaur et al., 2019)	Obtained from our collected suicide-related posts
Domain Knowledge	O	O	X
Size	2,168 words	7,286 sentences	1,000 words
Example	*friend, parents* *sorry, mental illness* *miss, god, uneasy*	*The present also makes me sad.* *Nothing can change now.* *God this seems pathetic.*	*suicide, live, want to die,* *father, mother, depression* *pain, grade, dream*

Table 2: A summary of the suicide dictionaries for training the suicidal-oriented word embeddings.

4 Suicide Data

To develop models for predicting suicidal ideation for a post written in Korean, we collected the suicide-related and non-suicide-related Korean posts from *Naver Cafe*[1]. To improve the model performance, we further collected data for generating suicide word embeddings for Chinese, English, and Korean, respectively. Note that all the collected data is anonymized, hence no user information can be identifiable.

4.1 Target Post Data for Predicting Suicidal Ideation

To collect suicide-related and non-suicide-related posts, we selected *Naver Cafe* operated by *Naver*, one of the most popular web-based services in Korea (Nam et al., 2009; Park et al., 2014). Like a subreddit in Reddit, a user can create a topic-based community in Naver Cafe, named a 'cafe', where members in a cafe can communicate with others via writing posts or commenting posts.

We collected 10,000 suicide-related posts from a representative suicide-related cafe in Korea, *'Talking about Suicide'*, where users share their interest in suicide, and 21,723 non-suicide-related posts from two popular cafes, *'Goodbye Single*[2]*'* and *'Cafe Powder Room*[3]*'*, where people socialize with others and share their daily life. Table 1 shows the examples of the suicide-related and non-suicide-related posts.

4.2 Data for Suicidal-Oriented Word Embedding

To train the suicidal-oriented word embeddings for each language (i.e., Korean, English, and Chinese), we further collected three language sets of data to generate suicidal and non-suicidal expressions, explained in Section 3.1

4.2.1 Suicide Dictionary

We first obtained the pre-built existing suicide dictionaries based on domain knowledge in Chinese and English. We also created a suicide dictionary in Korean to evaluate the model performance with this dictionary, written in the same language (i.e., Korean) of our target suicide-related posts but is computationally generated without any medical knowledge base. We detail the suicide dictionary for each language, summarized in Table 2, as follows.

A **Chinese suicide dictionary** is built by Lv et al. (2015), which includes 2,168 words extracted from 1.06 M posts in Sina Weibo. Note that each word has a score in the range of 1 to 3, assigned by three experts, indicating how strongly the given the word expresses suicidal ideation.

We obtained an **English suicide dictionary**, titled as *Gold Standard Dataset*, which was developed by Gaur et al. (2019). It contains 500 users' posts in the "r\SuicideWatch" subreddit in Reddit. Each user is annotated with one of the five levels across suicide severity (i.e., Indicator, Ideation, Behavior, Attempt, and Supportive) by practicing psychiatrists. We used only four levels except for the 'supportive' level to avoid confusion because

[1] http://cafe.naver.com/

[2] https://cafe.naver.com/dohak27

[3] https://cafe.naver.com/cosmania

a user in the 'supportive' class can be regarded as one without suicide risk but show similar linguistic characteristics to the users in other classes. Finally, we obtained 7,286 posts written by 373 users.

To generate a **Korean suicide dictionary**, we collected posts from representative suicide-related online communities in Korea. We collected 1,258 and 6,332 suicide-related posts from two suicide-related public web forums, *"Lifeline Korea"*[4] and *"Companions of Life, Suicide Prevention Counselling"*[5], where a user can share his/her suicidal ideation and can be supported by a mental health counselor. We further collected 2,410 suicide-related posts from the Naver cafe, *'Talking about Suicide'*. Note that additional data collected from the Naver cafe is only used for generating the Korean dictionary and not used for learning for detecting suicidal ideation. Following the method proposed in prior work (De Choudhury et al., 2013; Burnap et al., 2015), we then extracted the top 1,000 keywords from all the collected posts using the TF-IDF.

4.2.2 Posts for Word Embedding

To train a suicidal-oriented word embedding, we further collected suicide-related posts for Chinese and Korean, which use the suicide dictionary with word-level information, and non-suicide-related posts for English, which use the dictionary with sentence-level information, for generating suicidal and non-suicidal expressions. In particular, for training the word embedding for Chinese, we collected the *Tree Hole* posts in Sina Weibo, used in prior study (Cao et al., 2019; Zhao et al., 2018), where users have shared their thoughts on suicide by exchanging over 100 M comments. By using the Weibo API, we obtained 6,093 posts from March 11th to 31st in 2020. For training the word embedding for Korean, we obtained another set of the 2,410 suicide-related posts from the Naver cafe, *'Talking about Suicide'*, where we collected data for predicting suicidal ideation. Note that the newly added data is only used for word embedding. For training the English word embedding, we collected the 102 K non-suicide-related posts from the three subreddits in Reddit, "r\AskReddit", "r\Showerthoughts", and "r\CasualConversation", where users share casual topics or daily events.

4.3 Language Difference on Suicide

To analyze whether and how similar suicide topics are shared across different languages, we compare the top 100 keywords identified in each language. To identify the suicide-related keywords, we further collected 107,606 English suicide-related posts from the "r\SuicideWatch" subreddit in Reddit and 6,297 Chinese non-suicide-related posts from the "Popular" section in Weibo. Note that non-suicide-related posts were used to exclude the generally popular keywords from the top suicide-related keywords. Table 3 summarizes the numbers of the posts for each language, respectively, used in this analysis.

Language	Suicide-related post	Non-suicide-related post
Korean	17,351	27,788
Chinese	6,093	6,297
English	107,606	101,298

Table 3: The number of suicide-related and non-suicide-related posts for analysis on linguistics difference of suicide between Chinese, English, and Korean.

To compare the different languages, Korean and Chinese posts were translated into English by using the Naver Papago (Lee et al., 2016). We then performed the stemming and extracted unigrams and bigrams. To exclude the commonly used keywords in both suicide-related and non-suicide-related posts, we removed the keywords that also appear in the top 100 keywords for the non-related posts. Finally, we obtained the top 100 keywords from the suicide-related posts in each language.

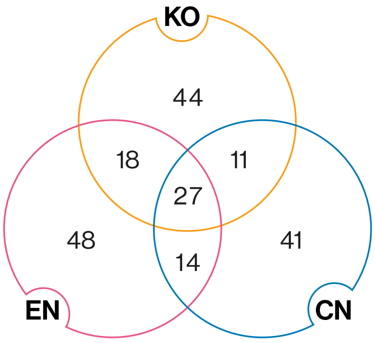

Figure 4: A Venn diagram illustrating how the identified top 100 keywords for the suicide-related posts in different languages are overlapped.

Figure 4 shows a Venn diagram that represents how the identified top 100 keywords for the suicide-related posts in different languages are overlapped. As shown in Figure 4, the 27 keywords are commonly identified in all the languages. In particular,

[4] https://www.lifeline.or.kr/
[5] http://www.counselling.or.kr/

the commonly overlapped words tend to directly express suicidal ideation (*'die', 'want die'*), show negative emotion (*'hate', 'cry', 'hurt', 'sad', 'wrong', 'pain'*), and mention about family (*'mother', 'mom', 'parents', 'family', 'dad'*). The intersection between Korean and English (45 words) includes more common keywords than that between Korean and Chinese (38 words) or between Chinese and English (41 words). This implies that Korean and English tend to share common topics on suicide than others more. For example, we find that the overlapped keywords for Korean and English tend to be related to loneliness (*'left', 'alone'*) and hope (*'able', 'want'*).

Taking a close look at the unique 44 top keywords in Korean, which is the target language for our model evaluation in Section 4.1, we find that the keywords tend to mention about life plan (*'job', 'dream'*), school life (*'high school', 'middle school', 'student', 'grade'*), beauty (*'face'*), sibling (*'brother', 'sister'*), and past (*'year ago', 'ago'*), which are not observed in other languages, i.e., English and Chinese.

In summary, our analysis reveals that utilizing the suicide word embedding for other languages can help improve the performance of our model that predicts suicide ideation, as different languages are likely to share similar topics. In addition, the ensemble of multiple languages in our model can be useful since it can capture the linguistic or cultural differences in suicide.

5 Experiments

We evaluate the proposed cross-lingual suicidal ideation detection model by answering the following research questions:

- RQ1: Can the word embedding refined by the suicide dictionary with domain knowledge in other foreign languages (e.g., Chinese, English) improve the model performance?
- RQ2: Is the refined word embedding based on the suicide dictionary created with a computational approach (without domain expert knowledge) useful in identifying suicidal ideation, compared to one with a pre-built existing suicide dictionary created by domain experts in a foreign language?
- RQ3: Can an ensemble from the multiple models with different languages improve the model performance?

5.1 Models

To answer the above questions, we evaluate the following models:

- ML-baseline is the mono-lingual (ML) model, which takes a post written in a single language as an input. For example, ML-baseline (language: CN) indicates a model taking a post written in Chinese translated from Korean as an input. Note that we use the well-known general pre-trained embeddings, i.e., Word2vec (Le and Mikolov, 2014) and FastText (Joulin et al., 2017).
- ML-refined is the same as the ML-baseline but uses the word embedding refined by the suicide dictionary, as explained in Section 3. For example, ML-refined (language: English, word-embedding: refined-word2vec) represents a model that uses the word2vec word embedding refined by the English suicide dictionary to learn posts written in English translated from Korean.
- CL-mixed is an ensemble cross-lingual (CL) model that combines multiple mono-lingual language models, e.g., ML-baseline (Korean) and ML-refined (Chinese). Note that we use the general pre-trained word embedding (e.g., word2vec) for the language where the suicide dictionary with domain knowledge does not exist (i.e., Korean), and the refined word embedding(s) for the language(s) where the suicide dictionary is constructed by domain experts (i.e., Chinese and English).
- CL-ours is the same as the CL-mixed but uses the word embedding refined by the suicide dictionary for the input language, Korean.

5.2 Results

To answer the questions, we evaluate the performance of each model, summarized in Table 4. Note that we conduct a 5-fold cross-validation. Our model uses the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.001 and epoch size of 30.

5.2.1 RQ1: Effect on Using Suicidal-Oriented Word Embedding in Other Language

To answer the first research question on using suicidal-oriented word embedding in other languages such as English or Chinese, we compare the

Model	Language	Word Embedding	Accuracy	Precision	Recall	F1-score
ML-baseline	KO	Word2vec	85.74	86.15	85.24	85.67
		FastText	84.39	84.62	84.28	84.36
	CN	Word2vec	81.6	81.67	81.66	81.62
		FastText	80.85	81.00	81.34	80.88
	EN	Word2vec	82.79	83.21	82.2	82.68
		FastText	83.15	85.17	80.43	82.66
ML-refined	KO	Refined-word2vec	86.37	87.22	85.22	86.2
		Refined-fastText	85.85	86.32	85.26	85.76
	CN	Refined-word2vec	79.56	80.21	79.49	79.46
		Refined-fastText	80.58	80.58	80.87	80.63
	EN	Refined-word2vec	82.76	83.84	81.21	82.48
		Refined-fastText	82.75	84.78	79.94	82.24
CL-mixed	KO + CN	Word2vec + Refined-word2vec	86.45	87.52	85.12	86.24
		FastText + Refined-fastText	86.03	86.02	86.05	86.02
	KO + EN	Word2vec + Refined-word2vec	86.84	87.77	85.71	86.68
		FastText + Refined-fastText	85.63	87.5	83.58	85.28
	KO + CN + EN	Word2vec + Refined-word2vec	86.89	88.04	85.64	86.72
		FastText + Refined-fastText	86.51	**89.43**	82.81	85.99
CL-ours	KO + CN	Refined-word2vec	87.04	87.28	86.77	86.99
		Refined-fastText	86.16	87.57	84.5	85.9
	KO + EN	Refined-word2vec	86.94	87.21	86.63	86.88
		Refined-fastText	86.64	86.35	87.16	86.71
	KO + CN + EN	Refined-word2vec	**87.50**	87.57	**87.41**	**87.49**
		Refined-fastText	86.53	87.11	86.25	86.48

Table 4: Performance results of the proposed models.

results of ML-baseline and ML-refined models. We find that ML-refined (CN or EN) models show similar or lower performance than ML-refined (KR) or ML-baseline (KR) models, meaning that using an existing suicide dictionary developed by domain experts in other language does not help to improve the model performance. This may be due to the cultural difference in suicide-related languages, which was discussed in Section 4.3 that showed different language usage in suicide across different languages, e.g., the overlapped portion of suicidal-related keywords used both in Korean and Chinese is just 45%. Note that the ML-baseline (CN, EN) models perform lower than the ML-baseline (KR), indicating that the translated language (e.g., from Korean to Chinese) can be used in identifying suicidal ideation but shows a limited performance.

5.2.2 RQ2: Effect on Using Suicidal-Oriented Word Embedding Created by a Computational Approach

To answer the second question, we evaluate the model with suicidal-oriented word embedding created by a computation approach (without domain expert knowledge), the ML-refined (KO). As shown in Table 4, the performance of the ML-refined (KO) model is improved compared to the ML-baseline (KO) model. This implies that if a suicide dictionary generated by domain experts does not exist, a suicidal-oriented word embedding generated by a computational approach is useful in identifying suicidal ideation. This is because suicidal people tend to use their own special words (Gaur et al., 2019), and the computational approach can capture such distinct patterns.

5.2.3 RQ3: Effect on Using Cross-Lingual Suicidal-Oriented Word Embeddings

To evaluate our ensemble approach that uses multiple cross-lingual languages together, we compare CL-mixed (KO+CN) and ML-baseline (KO) models. As shown in Table 4, overall the CL-mixed models outperform ML-baseline models, meaning that our cross-lingual approach is useful in identifying suicidal ideation. By combining the model for Korean with one for Chinese or English, we find that a potential limitation due to the cultural language difference can be mitigated.

Lastly, the final model, CL-ours, shows the best performance that achieves 87.5% accuracy. This demonstrates that our proposed cross-lingual model can detect suicidal ideation with high accuracy, which has an important implication on preventing and managing possible suicide risks.

6 Conclusion

This paper proposed a cross-lingual suicidal ideation detection model that provides a cost-effective way to predict suicidal ideation with social media data written in a language where no suicide dictionary exists. We proposed to apply (i) suicidal-oriented word embeddings developed for other languages (i.e., English and Chinese), (ii) attention mechanism for post representation, and (iii) an ensemble approach to reflect potential cultural and language difference. The proposed model achieved high accuracy, over 87%, signifying its great utility in detecting suicidal ideation using social media data for preventing potential suicide risk in an early stage.

Acknowledgments

This research was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A1A02085647) and the MSIT (Ministry of Science and ICT), Korea, under the ICAN (ICT Challenge and Advanced Network of HRD) program (2020-0-01816) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation).

References

Pete Burnap, Walter Colombo, and Jonathan Scourfield. 2015. Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM on hypertext & social media*, pages 75–84.

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.

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.

Glen Coppersmith, Ryan Leary, Eric Whyne, and Tony Wood. 2015. Quantifying suicidal ideation via language usage on social media. In *Proceedings of the Joint Statistics Meetings, Statistical Computing Section*.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the 7th International AAAI conference on weblogs and social media*.

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.

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.

Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. 2019. Suicidal ideation detection: A review of machine learning methods and applications. *arXiv preprint arXiv:1910.12611*.

Armand Joulin, Édouard Grave, Piotr Bojanowski, and Tomáš Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 427–431.

Jina Kim, Jieon Lee, Eunil Park, and Jinyoung Han. 2020. A deep learning model for detecting mental illness from user content on social media. *Scientific Reports*, 10(1):1–6.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury. 2015. Detecting changes in suicide content manifested in social media following celebrity suicides. In *Proceedings of the 26th ACM on hypertext & social media*, pages 85–94.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*, pages 1188–1196.

Hyoung-Gyu Lee, Jun-Seok Kim, Joong-Hwi Shin, Jaesong Lee, Ying-Xiu Quan, and Young-Seob Jeong. 2016. papago: A machine translation service with word sense disambiguation and currency conversion. In *Proceedings of the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 185–188.

Bill Yuchen Lin, Frank F Xu, Kenny Zhu, and Seung-won Hwang. 2018. Mining cross-cultural differences and similarities in social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 709–719.

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.

Michael J McCarthy. 2010. Internet monitoring of suicide risk in the population. *Journal of affective disorders*, 122(3):277–279.

Kevin Kyung Nam, Mark S Ackerman, and Lada A Adamic. 2009. Questions in, knowledge in? a study of naver's question answering community. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 779–788.

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.

OECD. 2020. Suicide rates (accessed: May 23, 2020). https://data.oecd.org/healthstat/suicide-rates.htm.

Sangkeun Park, Yongsung Kim, Uichin Lee, and Mark Ackerman. 2014. Understanding localness of knowledge sharing: a study of naver kin "here". In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices services*, pages 13–22.

Umashanthi Pavalanathan and Munmun De Choudhury. 2015. Identity management and mental health discourse in social media. In *Proceedings of the 24th International Conference on World Wide Web*, pages 315–321.

Fuji Ren, Xin Kang, and Changqin Quan. 2015. Examining accumulated emotional traits in suicide blogs with an emotion topic model. *IEEE journal of biomedical and health informatics*, 20(5):1384–1396.

Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety through reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 58–65.

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.

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.

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.

Yoshihiko Suhara, Yinzhan Xu, and Alex'Sandy' Pentland. 2017. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In *Proceedings of the 26th International Conference on World Wide Web*, pages 715–724.

Xiaoli Zhao, Shaofu Lin, and Zhisheng Huang. 2018. Text classification of micro-blog's" tree hole" based on convolutional neural network. In *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*, pages 1–5.