Suicidal Risk Detection for Military Personnel

Sungjoon Park* [1,2], **Kiwoong Park*** [1], **Jaimeen Ahn** [1] **Alice Oh** [1]
[1] School of Computing, KAIST, Republic of Korea
[2] Upstage AI Research, Upstage, Republic of Korea
{sungjoon.park, marspak, jaimeen01}@kaist.ac.kr
alice.oh@kaist.edu

Abstract

We analyze social media for detecting the suicidal risk of military personnel, which is especially crucial for countries with compulsory military service such as the Republic of Korea. From a widely-used Korean social Q&A site, we collect posts containing military-relevant content written by active-duty military personnel. We then annotate the posts with two groups of experts: military experts and mental health experts. Our dataset includes 2,791 posts with 13,955 corresponding expert annotations of suicidal risk levels, and this dataset is available to researchers who consent to research ethics agreement. Using various fine-tuned state-of-the-art language models, we predict the level of suicide risk, reaching .88 F1 score for classifying the risks.

1 Introduction

Suicide is one of the major causes of death in the military. In some countries where military service is compulsory because of a conscription system, active-duty military personnel live in physical separation from their family and friends for an extended period of time, often against their will. In the Republic of Korea, for example, most men have the obligation to serve in the military for about a year and half, leading to a large population of about 600,000 in active duty as of this year. Many of them experience difficulty adapting to the isolated environment, and some of them are at risk of suicide.

One approach to detect the suicide risk signs of active-duty soldiers is the analysis of social media posts, similar to the approach used for detecting suicide risk of the general public (Milne et al., 2016; Yates et al., 2017; Zirikly et al., 2019). However, research on military suicide finds that there are distinct risk factors, such as combat exposure, injury, bereavement, and negative unit climate associated only with military service (Nock et al., 2013). For this reason, we cannot directly apply the findings of suicide risk research of the general public to the military personnel. In this paper, we take on the challenge of collecting social media posts related to military service in the Republic of Korea, annotating and analyzing them using NLP methods for detecting suicide risk of active-duty soldiers.

The first and most challenging step is to create an annotated dataset of military-related social media posts written by active-duty soldiers. We collect posts from a popular social question and answering (Q&A) platform. Anonymous posts are allowed, so we find that there is a considerable amount of military-related posts that contain possible suicide risk and other mental health issues. Annotation poses a challenge, as the mental health and suicide risk of soldiers should be analyzed by mental health experts experienced in the military setting. It is difficult, though, to find such experts, so we reached out to two separate groups of experts, military experts and mental health experts. We asked both groups for annotation, and our analysis includes the results of the annotation, as well as the results of the prediction of suicide risk.

Our focused contribution is in building a dataset of 2,791 social media posts written by military personnel in Korean with corresponding 13,955 expert annotations of suicidal risk levels. We fine-tune various state-of-the-art language models to classify the risks for developing simple yet effective baselines, achieving up to .88 F1 score.

2 Constructing Annotated Dataset

We describe the steps in collecting relevant posts, preprocessing, and annotations. We also explain

* These authors contributed equally.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2523–2531,
November 16–20, 2020.

Risk Level	Description
Imminent Risk (3)	• Expressing to self-harm or suicide directly and explicitly. • Making concrete plans for suicide: seeking access to hazardous tools or pills • Existence of triggering events, make a will, etc.
High Risk (2)	• Expressing to self-harm or suicide indirectly and implicitly • Desire for suicidal behavior, suicidal ideation, self-harming • Risk factors becoming severe due to stressful events, relationship problems, etc.
Low Risk (1)	• Expressing depressed, stressed, anxiety due to environmental or internal factors • Maladaptation to (military) service, but would become adaptable through adequate measures • Requiring continuous treatment with the therapist or psychiatrist
No Risk (0)	• No help required due to no risky sign detected • Expressing mild sadness • Simple questions

Table 1: Risk annotation criteria. All posts are annotated with risk leve lof 0 to 3. *Imminent Risk (3)*: the writer of the post needs urgent help. The post might show concrete plans to commit suicide, or triggering events. *High Risk (2)*: the writer needs attention, reporting and help. The writer expresses desire and thinking about suicide, and/or self-harming behaviors. *Low Risk (1)*: the writer needs help but not urgently. *No Risk (0)*: no risk of suicide.

how we dealt with research ethics concerns.

2.1 Data Collection

Collecting Posts. We collect relevant posts from Naver Knowledge iN, an online Korean Q&A (Question and Answering) platform in 2019. Like Quora.com, people ask questions through anonymous posts without length constraints. In some cases, users disclose their personal matters to obtain advice from others. To collect the relevant posts, we use 58 military-related keywords plus suicide or self-harm related terms. For instance, we use 'military force', 'army + self-harm', 'army + suicide', and so on. For every keyword, we collect the most recent 1,000 posts without any meta-data such as username and timestamp because these features could make person identification quite easy. Through this process, we collect 44,108 posts.

Preprocessing. We preprocess the collected posts in three steps. Since 58 search keywords return many duplicated posts, so we first remove duplicates which reduces the data size.

Then we manually remove any post that is written by family or friends of the soldier, or by anyone unrelated to military. We retain only posts written by soldiers themselves so that the trained model can detect suicide risk signals of the active military personnel based on their first-person account.

Next, we manually remove personally identifiable information and all named entities in the text. We found 44 unit names, 7 school names, 10 number of grades, 2 region names, and 2 personal names and user ids and replaced them with unidentifiable placeholders. After preprocessing, we are left with 2,791 posts. The average length of a post is 92.7 words.

Ethical Concerns. We carefully consider any potential ethical concerns with the entire process of this research. We collect posts only if they are publicly available from Naver Knowledge iN, and we do not collect any metadata of the posts because they can potentially be used to identify authorship. Also, we manually inspect every post to remove personally identifiable information, masking all named entities. These processes are costly and make it very difficult to build a large-scale corpus. Annotators are shown only anonymized posts, and annotated data will be available to researchers with express consent not to contact or de-anonymize any of the posts.[1] This study is reviewed and approved by the KAIST Institutional Review Board (#KH2019-122).

2.2 Annotating Suicidal Risk

5 annotators evaluated the degree of suicidal risk of writers (military personnel) in the anonymized posts. We annotate the risk at the *post-level* because anonymous posts do not have user names from the start, and the other posts are anonymized by removing user names due to the ethical concerns.

Annotation Criteria. Table. 1 shows our annotation criteria, which came from existing shared task settings (Milne et al., 2016; Zirikly et al., 2019) and guidelines such as 'Classification criteria of *soldiers in need*' issued by the Ministry of National Defense. All posts are annotated with the risk level from 0 to 3, from lowest to highest level of detected

[1] https://github.com/SungjoonPark/RiskDetection

	E1	E2	I1	I2	I3
0	2,453	2,157	1,213	881	1,255
1	205	542	1,132	1,592	1,434
2	93	79	434	242	71
3	40	13	12	76	31

Table 2: Risk level distributions among annotators. External annotator group tends to evaluate most posts as *No Risk*, and internal annotator group labels more posts as *Low Risk*. The proportion of posts labeled as class *High Risk* and *Imminent Risk* is about 3-10%.

	E1	E2	I1	I2	I3
E1	1.00				
E2	0.54	1.00			
I1	0.22	0.33	1.00		
I2	0.19	0.24	0.45	1.00	
I3	0.25	0.36	0.45	0.55	1.00

Table 3: Cohen's inter-annotator Agreement (IAA) coefficients across annotators. The overall IAA shows fair agreement. Agreement within each group is higher than that of between groups.

suicidal risk.

Annotation Perspective. Evaluating suicidal risk of military personnel requires professional clinical knowledge as well as military experience because of various factors unique for the military setting (Nock et al., 2013; Oh and Lee, 2017). So we separate and consider the two perspectives: first of military experts and second of clinical experts. Military experts evaluate the risk as an insider of the special population, since they are familiar with military situational factors through their experiences living and working with 'soldiers in need'. We refer to it as the *internal* perspective. On the other hand, clinicians would view the posts and patients from outside of military service with their clinical experience and knowledge. So we refer to it as the *external* perspective. The difference of perspective for each expert group and detailed annotated examples are in Table. 4.

Annotation Process. We recruit two *external* expert annotators (E1 (Psychiatrist), E2 (Psychotherapist)) and three *internal* expert annotators (I1 (Military Counselor), I2 (Commander), I3 (Commander)). Each annotator independently evaluated the level of the risk detected in the 2,791 posts into the 4 classes. With posts that show disagreement within each group, annotators were asked to evaluate the risk of those items once again independently. Annotators could choose whether to change their initial evaluation, but in most cases they did not change the first evaluation.

2.3 Annotation Results

Here we show risk level frequencies for each annotator, and degrees of agreement among them.

Distribution of Risk Levels. As shown in Table. 2, most posts are labeled as *No Risk* or *Low Risk*. External annotators tend to evaluate most posts as *No Risk*, and internal annotators label more posts as *Low Risk* rather than *No Risk*. The proportion of posts labeled as *High Risk* and *Imminent Risk* is about 3-10%, showing skewed frequency distribution among classes.

Inter-Annotator Agreement (IAA). Table. 3 presents Cohen's IAA among annotators. We find the agreement of within group is higher than that of between groups.

In detail, within each group, the external annotators show fair agreement (Krippendorff's α=0.58), and the internal annotators fair agreement as well. (Krippendorff's α=0.55) The overall agreement among the five annotators is lower (Krippendorff's α=0.37) than that of within group agreements. In addition, if annotations are binarized to 'Flagged', the Krippendorff's αs are internal's α=0.53, external's α=0.52, overall α=0.30. Again, we observe fair agreement within groups, but the level of agreement of all annotators is rather low because of the difference between groups.

Comparison between Perspectives. Table. 4 shows a few manually selected examples of risk annotation between groups. For the first and second examples, two groups annotated the same scores. The first example explicitly expresses suicidal thoughts and even a failed suicide attempt, so all annotators agreed the writer seems to require immediate help. The second example asks merely for a skin problem, and both groups annotated as *No Risk* for this post.

Through third and fourth examples, we can see the difference between the two perspectives. For the third example, the internal annotators rated relatively higher than the external annotators. This is because the writer is under stress due to the problems adjusting to military life, suffering from depression and thinking about suicide. Also, they think this would have highly negative effect on the poster's life in the unit, so they judge the risk factor is relatively high, and the military must pay attention from the perspective of the commander responsible for this soldier's life and work in the military.

Annotation	Example
E (High, 3.00) I (High, 3.00)	I'm a soldier now.. I'm so tired of depression, insomnia, and hallucination. Every day I try to sleep, some voices tell me not to sleep, so I can't sleep without medicine... When I'm in a group of people, my heart beats fast, hard to breathe, head hurts, and I feel dizzy. And when I look in the mirror, I'm so surprised to see someone behind me, even though nobody's there. It was a shock to me... because I can definitely hear and see it. I feel like I'm lying. I don't know what's happening. I feel so sorry to my family, but after I die, hard time would be just a moment. I heard that 'Actifed' is bad for people with high blood pressure like me. I bought 100 pills at the pharmacy. I took all tablets of Actifed, but just throwing up 4 times and being paralyzed for an hour. I still feel pain when I move. I'm sorry I couldn't die.
E (No, 0.00) I (No, 0.00)	Although I'm in the military, I can go on vacation more often so I can take care of my skin. So far, I haven't cared much about my skin but then.. Should I take care of it since I'm on duty? Or after being discharged?
E (Low, 0.50) I (High, 2.00)	After a few days I was assigned to my platoon, I have felt weird symptoms. Due to the rebuke and curse from senior soldiers, I was so nervous that I couldn't carry out my mission efficiently, and couldn't think or judge well just like a teenager. I hate to be with others, and I didn't really want to live every day. I usually come up with suicidal ideas, but I try hard to withstand the situation by thinking of my parents, and shedding tears alone. I have a continuous headache with dizziness, get to sleep irregularly, and I'm in a daze. It feels like what I'm doing isn't mine, and I feel depressed all day long. I don't even have an appetite. But unit says it's difficult to discharge me early because my situation is not bad enough, and it doesn't look very serious. My unit's refusing though my medical report says "Consider maladaptation to service." I'm having a hard time every day. I want to get out of the unit and get counseling and proceed with treatment.
E (High, 2.00) I (Low, 0.67)	I wanna kill myself How can I stand military life for 9 more months.. I could be reminiscing looking back on this later. I don't know if I can do it well

Table 4: Examples of risk annotation between groups. The first and second examples show the same evaluation of the risk for the post. For the third and fourth examples, there is disagreement of the evaluated risk between the two expert groups.

However, the external annotators expect that there is little suicidal risk because the writer wants to visit a therapist or psychiatrists anyway rather than moving onto suicidal behavior. The fourth example expresses thoughts about killing oneself, so the external annotators give a higher risk score to this example. But the internal annotators see less risk from the post because they have commonly heard this type of negative expression about the mandatory military service among the soldiers.

Considering these examples and others, we conclude that both perspectives should be considered together and separately while predicting the suicidal risk of military personnel using a computational model.

3 Experiments

We classify the posts by annotated risk levels at the post-level. We first compute the *maximum* value of risk annotations to aggregate multiple annotations for each post with the aim to give alert if there is *any* possibility of suicidal risk. This might increase false positives, but the experts view that in practice, false positives are better than false negatives.

Classification Types. We classify posts in three ways: 1) the four risk levels, 2) `Flagged` or not, and 3) `Urgent` or not. For binary classification, we consider *Low, High, Imminent Risk* posts as `Flagged` and *No Risk* as `Not Flagged`, and *High, Imminent Risk* as `Urgent` and *Low, No Risk* posts as `Not Urgent` (Milne et al., 2016).

Models. We leverage two type of models: 1) Convolutional Neural Network (CNN) and 2) pre-trained language models, which are used in the relevant shared task (Zirikly et al., 2019). The teams that participated in the shared task demonstrated that CNN is effective for the risk classification task (Morales et al., 2019). Also, ASU (Ambalavanan et al., 2019) shows fine-tuning pre-trained language model is highly effective. Note that our dataset contains Korean posts with post-level risk annotations, so these previous models should be adjusted to our dataset.

Specifically, we use CNN with pre-trained Korean subword-level word embeddings for the input of the two convolution layers (Park et al., 2018). In case of using pre-trained language models, we have a choice to use *multilingual* models trained

Model	All annotators			Internal expert			External expert		
	4-level F1 (Acc.)	Flagged F1 (Acc.)	Urgent F1 (Acc.)	4-level F1 (Acc.)	Flagged F1 (Acc.)	Urgent F1 (Acc.)	4-level F1 (Acc.)	Flagged F1 (Acc.)	Urgent F1 (Acc.)
CNN	0.56 (0.71)	0.87 (0.90)	0.76 (0.82)	0.52 (0.72)	0.88 (0.91)	0.76 (0.84)	0.46 (0.81)	0.81 (0.86)	0.72 (0.94)
BERT-Multilingual	0.65 (0.72)	0.85 (0.89)	0.74 (0.84)	0.63 (0.71)	0.84 (0.89)	0.71 (0.84)	0.51 (0.84)	0.82 (0.87)	0.81 (0.96)
KoBERT	**0.72** (0.76)	**0.88** (0.91)	**0.80** (0.86)	**0.68** (0.75)	**0.88** (0.92)	**0.80** (0.87)	0.55 (0.84)	**0.85** (0.88)	0.81 (0.96)
XLM-R	0.70 (0.75)	0.87 (0.90)	**0.80** (0.86)	**0.68** (0.75)	0.87 (0.90)	**0.80** (0.86)	**0.56** (0.85)	**0.85** (0.88)	**0.83** (0.96)

Table 5: Results of classification models. Fine-tuning BERT for Korean (KoBERT) or multilingual RoBERTa (XLM-R) shows better performance compared to BERT-multilingual and CNN. We emphasize that fine-tuned language models can correctly classify the post which needs *urgent* help with 0.80 F1 score and 86% accuracy.

with a corpus that includes Korean (Multilingual-BERT (Devlin et al., 2018), XLM-R (Conneau et al., 2020)) , or a model pre-trained over only Korean corpus (KoBERT [2]). Since we classify the risk at the post-level, these models are fine-tuned by post-level supervisions without aggregation of posts to the user-level. Detailed experimental settings for reproducibility are in the Appendix.

Results. The results are shown in Table. 5. We report accuracy as well as macro F1 to consider precision and recall on the test sets. Overall, KoBERT and XLM-R outperform the others. Corresponding results on the validation set and most frequent baselines are in Appendix.

For classifying posts with maximum risk levels from all annotators, KoBERT outperforms in classifying posts at the four risk levels (F1 = 0.72, acc = 0.76), and whether the post is `Flagged` or not. (F1 = 0.88, acc = 0.91) XLM-R shows comparable performance in classifying the `Urgent` posts. (F1 = 0.80, acc = 0.86) This tendency is shown in classifying the internal expert risk level annotations. For External expert's risk classification, XLM-R shows slightly better performance.

Among the annotator groups, internal expert group tends to obtain high scores in overall for both F1 and accuracy (F1 = 0.88, acc = 0.92), while external expert group shows the lowest F1 in average in 4-level risk classification. (F1 = 0.56, acc = 0.85) This is caused by the small number of *Imminent risk* in the external annotator's evaluations.

Also, `Flagged` posts classification performance is higher than that of 4-level or `Urgent` post classification, which implies our model identifies well the posts with any level of risk from posts with *No Risk*. In the 'Flagged' condition, label (1), (2) and (3) are combined as a single class, so imbalance between classes is partially relieved, which leads to a better F1 score. In practice, this would be quite helpful for consideration of intervention.

[2] https://github.com/SKTBrain/KoBERT

4 Discussion and Conclusions

In this paper, we tackle the problem of suicidal risk of military personnel from their social media posts. We focus on the specific population of military personnel in compulsory service because it requires a unique approach to fully understand their suicidal risk. As our first step, we collect 2,791 military-relevant posts in a social Q&A platform that are written by at-risk active-duty soldiers and remove any identifying information from the data. Then five annotators (three military experts and two clinicians) evaluate the degree of suicidal risk of the posts. After the dataset is constructed, we fine-tune a pre-trained language models, achieving at most 0.88 F1 score.

Our research can be the first step toward proper intervention programs and institutional support for soldiers with mental health issues. Such follow-up would maximize the value of our model and data. We also plan to add domain-specific features to our model, collect more data, integrate existing suicidal risk datasets with various languages to improve performance.

Acknowledgments

This research was supported by the Engineering Research Center Program through the National Research Foundation of Korea (NRF) funded by the Korean Government MSIT (NRF-2018R1A5A1059921)

References

Ashwin Karthik Ambalavanan, Pranjali Dileep Jagtap, Soumya Adhya, and Murthy Devarakonda. 2019. Using contextual representations for suicide risk assessment from Internet forums. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 172–176, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael D Anestis, Richard S Mohn, Jack W Dorminey, and Bradley A Green. 2019. Detecting potential underreporting of suicide ideation among us military personnel. *Suicide and Life-Threatening Behavior*, 49(1):210–220.

Mario Ezra Aragón, Adrian Pastor López-Monroy, Luis Carlos González-Gurrola, and Manuel Montes. 2019. Detecting depression in social media using fine-grained emotions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1481–1486.

Shrey Bagroy, Ponnurangam Kumaraguru, and Munmun De Choudhury. 2017. A social media based index of mental well-being in college campuses. In *Proceedings of the 2017 CHI Conference on Human factors in Computing Systems*, pages 1634–1646. ACM.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multi-task learning for mental health using social media text. *arXiv preprint arXiv:1712.03538*.

Craig J Bryan, Jonathan E Butner, Sungchoon Sinclair, Anna Belle O Bryan, Christina M Hesse, and Andree E Rose. 2018. Predictors of emerging suicide death among military personnel on social media networks. *Suicide and Life-Threatening Behavior*, 48(4):413–430.

Craig J Bryan, Ann Marie Hernandez, Sybil Allison, and Tracy Clemans. 2013a. Combat exposure and suicide risk in two samples of military personnel. *Journal of clinical psychology*, 69(1):64–77.

Craig J Bryan, Chad E Morrow, Neysa Etienne, and Bobbie Ray-Sannerud. 2013b. Guilt, shame, and suicidal ideation in a military outpatient clinical sample. *Depression and anxiety*, 30(1):55–60.

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, Hong Kong, China. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 2020 Annual Conference of the Association for Computational Linguistics (ACL)*.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

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

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Seventh 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. ACM.

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*.

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Christopher Homan, Ravdeep Johar, Tong Liu, Megan Lytle, Vincent Silenzio, and Cecilia Ovesdotter Alm. 2014. Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 107–117.

Jun Su Jung, Sung Jin Park, Eun Young Kim, Kyoung-Sae Na, Young Jae Kim, and Kwang Gi Kim. 2019. Prediction models for high risk of suicide in korean adolescents using machine learning techniques. *PLoS one*, 14(6):e0217639.

Sunah Kim, Hyun Lye Kim, Chunghee Woo, Suin Park, and Ran Keum. 2011. Communication abilities, interpersonal relationship, anxiety, and depression in korean soldiers. *Journal of Korean Academy of Psychiatric and Mental Health Nursing*, 20(1):81–90.

Raina M Merchant, David A Asch, Patrick Crutchley, Lyle H Ungar, Sharath C Guntuku, Johannes C Eichstaedt, Shawndra Hill, Kevin Padrez, Robert J Smith, and H Andrew Schwartz. 2019. Evaluating the predictability of medical conditions from social media posts. *PloS one*, 14(6):e0215476.

David N. Milne, Glen Pink, Ben Hachey, and Rafael A. Calvo. 2016. CLPsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Michelle Morales, Prajjalita Dey, Thomas Theisen, Danny Belitz, and Natalia Chernova. 2019. An investigation of deep learning systems for suicide risk assessment. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 177–181, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew K Nock, Charlene A Deming, Carol S Fullerton, Stephen E Gilman, Matthew Goldenberg, Ronald C Kessler, James E McCarroll, Katie A McLaughlin, Christopher Peterson, Michael Schoenbaum, et al. 2013. Suicide among soldiers: a review of psychosocial risk and protective factors. *Psychiatry: Interpersonal & Biological Processes*, 76(2):97–125.

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.

Dae Jong Oh Oh and Sang Don Lee. 2017. Review of the risk factors associated with suicide and suicide-related behavior in military personnel. *Journal of the Korean Society of Biological Therapies in Psychiatry*, 23(1):13–22.

Sungjoon Park, Jeongmin Byun, Sion Baek, Yongseok Cho, and Alice Oh. 2018. Subword-level word vector representations for korean. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2429–2438.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Mark A Reger, Raymond P Tucker, Sarah P Carter, and Brooke A Ammerman. 2018. Military deployments and suicide: a critical examination. *Perspectives on psychological science*, 13(6):688–699.

H Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 118–125.

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.

Amanda R Start, Yvonne Allard, Amy Adler, and Robin Toblin. 2019. Predicting suicide ideation in the military: the independent role of aggression. *Suicide and Life-Threatening Behavior*, 49(2):444–454.

Paul Thompson, Craig Bryan, and Chris Poulin. 2014. Predicting military and veteran suicide risk: Cultural aspects. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–6.

Chen-Kai Wang, Onkar Singh, Zhao-Li Tang, and Hong-Jie Dai. 2017. Using a recurrent neural network model for classification of tweets conveyed influenza-related information. In *Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017)*, pages 33–38.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*.

Ayah Zirikly, Philip Resnik, Özlem 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, Minneapolis, Minnesota. Association for Computational Linguistics.

A Experimental Details

A.1 Comparison Models

We compare the classification performance of comparison models trained on our dataset. We use accuracy and F1 score as evaluation metrics, computing them using scikit-learn. (Pedregosa et al., 2011)

Convolutional Neural Networks. (CNN) We use CNN for text classification. For the input of the convolutional layer, we use 300 dimension pre-trained Korean subword-level word embeddings (Park et al., 2018), and set the number of layers as 2, the number of filters as 256. On the top of the layers, we add sigmoid or softmax activation function for classification

Multilingual BERT. Since our dataset consists of Korean posts, we use Multilingual version of BERT-base cased model. (Devlin et al., 2018) We add a trainable linear layer and sigmoid or softmax activation over the classification head ([CLS]). Then the entire model is fine-tuned by minimizing the cross entropy loss between the predicted label and the target labels, using BertAdam optimizer (Devlin et al., 2018). Most hyperparameters are chosen in original paper. (Devlin et al., 2018) The number of trainable parameters is 110M.

KoBERT. A BERT-base model trained on a corpus consisting of Korean Wikipedia corpus and news data to improve the performance of the multilingual BERT. This is a pre-trained model which is publicly available in GitHub. We use this model to fine-tune on our data. We set all details the same as described in Multilingual BERT above. The number of trainable parameters is 110M.

XLM-R. A state-of-the-art pre-trained cross-lingual language model which trained on corpus including Korean Documents. (Conneau et al., 2020) The model is based on RoBERTa architecture, and the pre-trained model is shown to be highly effective cross-lingual language understanding tasks. Like other BERTs, we fine-tune this model on our data with adding the same classification layer. The number of trainable parameters is 275M.

A.2 Hyperparameters

Batch size is set to 32, and the maximum sequence length to 512 in CNN, Multiligual BERT, KoBERT, and XLM-R. The learning rate of all models are set to 3e-5 like previous settings. (Devlin et al., 2018; Conneau et al., 2020) The batch size and the sequence length is manually chosen to fit the models to our computing infrastructure. All models are trained on single RTX 2080Ti GPU. For every run, a model converged at most within 3 hours.

A.3 Data Splits

We train the classifiers on the training set which consists of 1,674 posts, and evaluated on the 559 posts of the test set. The number of examples in each splits are shown in Table. 7-9 We use a validation set which contains 558 posts for tuning the hyperparameters of our models.

A.4 Most Frequent Baselines

When aggregating all 5 annotators' labels in 4 risk levels, *Low Risk* label accounts for 54.28% of all posts. In binarized label as `Flagged` or not, `Flagged` is relatively more frequent (75.24%) than `Not flagged`, and another binarized label as Urgent or not, `Not urgent` accounts for 78.68%.

Aggregating labels of 3 internal experts shows similar tendency with all 5 annotators' result. `Low Risk` label is the most frequent class which accounts for 54.32% in 4 risk levels. `Flagged` posts are more frequent (75.42%) than `Not flagged`, and `Not urgent` posts are relatively more frequent(78.90%) than `Urgent` post.

2 External experts' annotations are quite different from other group's results. In 4 levels of risk, the ratio of *No Risk* posts reaches to 78.68%, which is the most frequent, and this ratio is obviously same in `Not flagged` posts. When dividing labels into `Urgent` or not, `Not urgent` occupies 94.66%.

Comparing to the most frequent class baseline, the results of our classification models are higher in terms of accuracy.

B Related Work

Detecting Mental Illness in Social Media. Social media posts are widely used in mental illness research using computational methods. One common approach is detecting mental illness related variables from the posts automatically, such as depression (De Choudhury et al., 2013; Schwartz et al., 2014; Guntuku et al., 2017; Eichstaedt et al., 2018), self-harm (Milne et al., 2016; Yates et al., 2017), and suicidal risk (Homan et al., 2014; O'Dea et al., 2015; De Choudhury et al., 2016; Coppersmith et al., 2018; Cao et al., 2019; Aragón et al., 2019; Jung et al., 2019). These approaches usually aim to help people in need immediately. More

Model	All annotators			Internal expert			External expert		
	4-level F1 (Acc.)	Flagged F1 (Acc.)	Urgent F1 (Acc.)	4-level F1 (Acc.)	Flagged F1 (Acc.)	Urgent F1 (Acc.)	4-level F1 (Acc.)	Flagged F1 (Acc.)	Urgent F1 (Acc.)
CNN	0.59 (0.72)	0.89 (0.93)	0.77 (0.82)	0.55 (0.74)	0.89 (0.93)	0.75 (0.76)	0.46 (0.80)	0.80 (0.85)	0.77 (0.95)
BERT-Multilingual	0.68 (0.75)	0.86 (0.91)	0.74 (0.85)	0.65 (0.73)	0.86 (0.91)	0.76 (0.85)	0.48 (0.82)	0.79 (0.85)	0.81 (0.96)
KoBERT	**0.75** (0.78)	**0.91** (0.94)	**0.81** (0.87)	**0.73** (0.78)	**0.90** (0.94)	**0.81** (0.87)	0.61 (0.84)	**0.85** (0.89)	0.86 (0.97)
XLM-R	0.73 (0.76)	0.90 (0.94)	**0.81** (0.87)	**0.70** (0.76)	0.91 (0.94)	**0.81** (0.86)	**0.57** (0.83)	**0.85** (0.89)	**0.85** (0.97)

Table 6: Results of classification models on validation set. F1 scores are bold if it is the highest in test set.

All	No Risk	Low Risk	High Risk	Immi. Risk
train	416	913	289	56
valid	119	312	105	22
test	146	290	104	19

Table 7: Number of examples in each train, valid, test splits when aggregating *All* experts' annotations.

Internal	No Risk	Low Risk	High Risk	Immi. Risk
train	420	914	293	47
valid	119	313	105	21
test	147	289	106	17

Table 8: Number of examples in each train, valid, test splits when aggregating *Internal* experts' annotations.

general mental health research includes predicting mental health conditions (Benton et al., 2017), mental well-being (Bagroy et al., 2017), physical illness (Wang et al., 2017), and medical conditions (Merchant et al., 2019).

This kind of research requires annotated data, so much effort has been made toward data collection and dissemination. The 2nd Workshop on Computational Linguistics and Clinical Psychology (CLPsych'15) introduced a shared task (Coppersmith et al., 2015) to identify depression and post-traumatic stress disorder (PTSD) users using a Twitter dataset. The shared task for CLPsych'19 introduced an assessment of suicide risk based on social media postings using data from Reddit to identify the four levels of risk (Zirikly et al., 2019). Yates et al. (2017) introduced a large-scale Reddit dataset containing 9,000 users with self-reported depression diagnoses, along with over 107,000 control users. Another research created a general Reddit dataset for the assessment of suicide risk via online postings (Shing et al., 2018).

Unlike previous studies, our work focus on a specific at-risk population. Suicidal risk of military personnel could more easily result in tragic consequences because of their easier access to firearms (Nock et al., 2013; Oh and Lee, 2017).

External	No Risk	Low Risk	High Risk	Immi. Risk
train	1,266	320	62	26
valid	423	104	21	10
test	422	107	22	8

Table 9: Number of examples in each train, valid, test splits when aggregating *External* experts' annotations.

Mental Health Problems of Military Personnel. Since mental health problems in military are different from those of the general population, they should be treated distinctly. Previous research in soldiers' mental health looks into patients with PTSD and other traumatic experiences. This line of research mainly investigates the patients' medical records, questionnaires, psychological measurement tools, interviews, or administrative data (Kim et al., 2011; Bryan et al., 2013a; Thompson et al., 2014; Bryan et al., 2013b; Reger et al., 2018; Anestis et al., 2019; Start et al., 2019).

A study using social media posts investigates the temporal changes in military personnel's posts during the year preceding their death, through content coding method and multilevel models (Bryan et al., 2018). This work focuses on explaining the factors of suicide from posts, rather than train a model to predict the risks from unseen data.

Our work applies a computational method to social media posts to predict suicidal risks from unseen posts without additional manual coding. This research opens up an important new direction in computational analysis of mental health in a special at-risk population.