Linguistic Features of Clients and Counselors for Early Detection of Mental Health Issues in Online Text-based Counseling

Kazuhiro Shidara[1], Hiroki Tanaka[1], Rumiko Asada[2], Kayo Higashiyama[3], Hiroyoshi Adachi[4], Daisuke Kanayama[4], Yukako Sakagami[4], Takashi Kudo[4] and Satoshi Nakamura[1]

***Abstract*— Online counseling is essential for overcoming mobility restrictions, schedule limitations, and mental health stigma. However, government counseling offices are being inundated with consultations for which non-mental health supports are targeted. Therefore, we aim to create a classification model that classifies whether the clients have mental health issues or other issues. We expect to support counselors by presenting the classification results. We conducted the first automatic detection of clients who might be suffering from mental health issues and used almost 1000 actual counseling sessions for our machine learning framework. We achieved an F1-score of 0.646 by classifying dialogue sessions using features such as frequency-inverse, document frequency, document embedding of a large-scale language model, linguistic inquiry and word count, topic modeling, and statistics of dialogue sentences. In addition, we performed dimensionality reduction with principal component analysis. We also conducted evaluation experiments using dialogue sentences from the beginning to the middle of sessions as input and clarified the relationship between the number of messages in the dialogues and the transition in the classification performance. We also identified the words that contribute to detecting mental health issues for each client and counselor.**

***Clinical relevance*—This study makes it possible to detect the trends identified in a client's anxieties during counseling. Our findings are critical for designing systems that assist counselors.**

I. INTRODUCTION

Text-based counseling using SNSs (online counseling) [1] is indispensable for overcoming travel restrictions, time limits, and psychological resistance. Since the recent spread of COVID-19, the need to respond to consultations due to various problems has increased [2]. The damage caused by this pandemic causes problems not only for the physical health of the infected counselors but also in the situations of their patients, such as mental health, work, family environment, and finance. At administrative counseling desks, counselors address such mental concerns as anxiety and depression. There are many other types of consultations than mental worries, such as physical conditions and work difficulties. Counselors need to address these issues, refer clients to appropriate support agencies, and provide professional care. Counselors need to differentiate whether the client needs counseling, financial support, medical institution guidance, and so on. However, what the client wants is often clarified as the counseling process itself progresses, and counselors are greatly burdened by the decision of whether to provide counseling to a client or to lead him to another window of support [3].

Therefore, it is necessary to provide technical support for the counselor's decision-making process of the care approach. We aim to create a system that mechanically classifies whether the client requires counseling or guidance to another window and presents the classification result to the counselor. A machine learning model that performs automatic classification based on previous counseling data is suitable for this system. Furthermore, by presenting to the counselor the features that influenced the detection of the classification model, the counselor can make a judgment based on the data. Therefore, in this paper, as a step toward building a counselor assistance system, we create a detection model for clients with mental health issues.

A task that resembles our study is the automatic detection of depressive tendencies. Previous studies have focused specifically on speech features [4], [5]. Multimodal detection methods that combine image and audio features have also been proposed [6], [7]. Many detection methods with text features use posts from social network services (e.g., Twitter) [8], [9]. Other similar studies attempt to classify counseling successes or failures using the text features of online counseling. Althoff et al. [10] and Xu et al. [11] used a counseling dataset to perform binary classification of successful or unsuccessful counseling based on counselor messages.

This study uses actual counseling data from counselors and clients to address the novel issue of support systems for counselors. We will not only build a classification model but also investigate which linguistic features strongly contribute to the classification. The interpretability of the judgment results is also critical in this classification model. In this study, we extracted multiple types of linguistic features and compared them using the same machine learning algorithm. By investigating how linguistic features are selected and how they relate to client/counselor interactions, we implemented a machine learning framework that can be applied to different counseling centers.

*This work was supported by JST CREST Grant Number JPMJCR19A5, Japan.

[1]Kazuhiro Shidara, Hiroki Tanaka, and Satoshi Nakamura are Nara Institute of Science and Technology, Japan {shidara.kazuhiro.sc5, hiroki-tan, s-nakamura}@is.naist.jp

[2]Rumiko Asada is with the Public Health and Medical Administration Office, Department of Public Health and Medical Affairs, Osaka Prefectural Government, Japan AsadaR@mbox.pref.osaka.lg.jp

[3]Kayo Higashiyama is with the Izumi Public Health Center, Osaka Prefectural Government, Japan HigashiyamaK@mbox.pref.osaka.lg.jp

[4]Hiroyoshi Adachi, Daisuke Kanayama, Yukako Sakagami, and Takashi Kudo are with Osaka University, Japan {hadachi, kanayama, kudo}@psy.med.osaka-u.ac.jp, sakagamiyukako@wellness.hss.osaka-u.ac.jp

TABLE I
NUMBER OF DIALOGUE SESSIONS FOR EACH TYPE OF CONSULTATION

Total number of issues		974
Mental health issues	Mental health	384
	Suicide thoughts	5
Other issues	Family	162
	Physical health	131
	Work	103
	Finance	81
	Sexual relationships	37
	Prejudice concerning COVID-19	5
	School	17
	Others	49

II. METHODS

A. Counseling Data

The dataset used in this study is the message records of actual counseling sessions between counselors and clients. The creator of this data set is a public interest foundation commissioned by the Osaka Prefectural Government, and the provider of this data set to the authors is the Osaka Prefectural Government. This dataset was anonymized before being provided to the authors. All efforts for this study were made with the approval of Osaka Prefecture. Its counseling platform is a messenger application called LINE (`https://line.me/`). The dataset was collected between May 2020 and January 2021. The labeling was done by one of the counselors actually doing the counseling. For completed counseling records, the counselor categorizes them by ten topics: mental health, suicide thoughts, family, physical health, work, finance, sexual relationships, prejudice concerning COVID-19, school, and others.

We divided those consultations into two categories, mental health issues (n=389) and other issues (n=585) under the supervision of the counselor who labeled them, another counselor who actually provided the counseling, and a psychiatrist. Table I shows the correspondence between these two categorizations and the ten different consultation categorizations labeled by the counselors and the number of consultations for each category. We subsumed the suicide-related issues under the mental health issues because they require mental health care. The counseling services targeted in this study specialize in such mental health issues as mental instability, depressive tendencies, and suicidal thoughts for which cognitive techniques are the primary method of intervention. On the other hand, there are specialized counseling centers outside of mental health for physical health issues and family concerns. Although mental health care can also be effective for these problems, most cannot be solved just through counseling. Therefore, a counseling service often directs them to a support service for each issue. Such counseling services are required to support and target specific mental health issues. We use our dataset to build a detection model for mental health issues.

B. Pre-processing and Feature Extraction

We used three training and test data settings: only client messages, counselor messages, and client and counselor messages. First, we conducted undersampling to resolve the imbalance among the classes and reduced the number of samples of other issues from 585 to 389, a number which is identical to the number of samples of mental health issues. MeCab (`http://taku910.github.io/mecab/`), a language analysis tool, was used for the tokenization and word normalization of the Japanese sentences. Our experiment used the following six feature settings from the sessions:

Term Frequency-Inverse Document Frequency (TF-IDF): The TF-IDF method considers the importance of words in a sentence to describe its features. The frequency of a word's use in a single document and its frequency of use in all the documents are used to calculate a word that represents each document. We calculated the TF-IDF using one dialogue session as one document, and the unit of words in the calculation is a unigram.

Latent Dirichlet Allocation (LDA): LDA is a topic model. Whereas TF-IDF calculated the surface meaning of a sentence based on the frequency of word occurrence, LDA calculates its latent topic [12]. In this study, we set the number of topics in LDA to 10, referring to the number of consultation categories in the dataset.

Contextualized document embeddings (Embeddings): For context-aware document embedding, we used a large-scale language model called Bidirectional Encoder Representations from Transformers (BERT) [13]. The input to BERT is a tokenized sentence, where a unique token called [CLS] is inserted at the beginning. We applied the [CLS] token encoding result, a 768-dimensional embedded representation, as a feature.

Linguistic inquiry and word count (LIWC): LIWC, a linguistic resource for categorizing words from a psychological perspective [14], is widely used in emotion analysis and the detection of mental disorders. We used LIWC, which was newly created for Japanese (`https://github.com/sociocom/JIWC-Dictionary`). Each word has a score for each of seven emotions: joy, trust, fear, surprise, sadness, disgust, and anger. We used the average scores of messages.

Statistics: We calculated the following descriptive statistics and used them as features: number of words (tokens), number of word types (types), type-token ratio (type/tokens), Guiraud's index (type/$\sqrt{tokens}$) [15], dialogue duration (seconds), number of messages, messages per seconds, and the ratio of the number of counselor and client messages. Table II shows mean values and results of the each descriptive statistics.

All features: In addition to individually using the features described above, we also conducted classification experiments when all the feature vectors were combined.

C. Dimensionality Reduction

Dimensionality reduction by principal component analysis was performed for all experimental settings. The principal components were extracted so that the cumulative contribution proportion was 70%.

TABLE II

MEAN FOR EACH LABEL AND UNPAIRED T-TEST RESULT: CL MEANS CLIENT, CO MEANS COUNSELOR. p VALUES LESS THAN 0.05 ARE SHOWN IN **BOLD**.

		Mental health issues	Other issues	p	Hedge's g (Effect size)
Tokens	Cl	950	879	0.061	0.117
	Co	959	868	**0.004**	0.191
Types	Cl	197	186	0.054	0.121
	Co	181	171	**0.008**	0.172
Types/Tokens	Cl	0.247	0.267	**<0.001**	-0.241
	Co	0.208	0.227	**<0.001**	-0.325
Types/$\sqrt{Tokens}$	Cl	6.55	6.53	0.706	0.024
	Co	5.96	6.00	0.439	-0.050
Seconds		5476	4848	**<0.001**	0.240
Number of messages	Cl	32.1	29.4	0.090	0.105
	Co	27.8	24.7	**<0.001**	0.219
Seconds /Messages	Cl	100	103	0.363	-0.057
	Co	100	104	0.217	-0.081
Ratio of the number of cl. and co. messages		1.11	1.20	0.377	-0.055

TABLE III

PERFORMANCE OF DETECTION MODELS: BEST PERFORMANCE IN EACH DATA SETTING IS SHOWN IN **BOLD**.

Data used for training and test	Feature	Precision	Recall	F1-score	AUC
Client	TF-IDF	**0.614**	0.681	**0.646**	**0.703**
	Embeddings	0.597	**0.698**	0.643	0.698
	LDA	0.393	0.600	0.475	0.503
	LIWC	0.391	0.500	0.439	0.501
	Statistics	0.393	0.450	0.420	0.502
	All features	0.452	0.479	0.465	0.553
Counselor	TF-IDF	**0.556**	**0.633**	**0.592**	**0.651**
	Embeddings	0.447	0.538	0.488	0.550
	LDA	0.401	0.503	0.446	0.505
	LIWC	0.399	0.500	0.444	0.503
	Statistics	0.414	0.428	0.421	0.515
	All features	0.448	0.459	0.453	0.543
Client and Counselor	TF-IDF	**0.576**	**0.668**	**0.618**	**0.678**
	Embeddings	0.552	0.661	0.602	0.661
	LDA	0.389	0.400	0.394	0.501
	LIWC	0.388	0.408	0.398	0.501
	Statistics	0.380	0.498	0.431	0.492
	All features	0.450	0.479	0.464	0.554

D. Logistic Regression Model

We used a logistic regression model. Since our study needs the interpretability of models as well as their higher classification performance, a logistic regression classifier is suitable because it provides high generalizability and interpretability. We used the selected features as the input of the classifiers, which were trained to predict mental health issues and other categories with the following default parameters. Binary classification was performed using the selected features as input. In the output label, 1 is a mental problem, and 0 is another problem. We used Scikit-learn (version 0.0) (`https://scikit-learn.org/stable/index.html`) to implement logistic regression. For the hyperparameters, regularization was L2 and the max_iter (the maximum number of searches when looking for optimal solutions) was 50×10^6. The other values are defaults. The solver is L-Broyden–Fletcher–Goldfarb–Shanno, and cost is 1.

E. Evaluation Metrics

We conducted stratified 10-fold cross-validation. This method generates test sets that all contain the identical distribution of classes or as close as possible. We reported the recall, precision, F1-score, and the area under a receiver-operating-characteristic curve (AUC). Next we analyzed the relationship between the F1-scores and the number of conversational messages as well as the important features to examine the interpretability of the results.

III. RESULTS AND DISCUSSIONS

A. Detection of Mental Health Issues

Table III shows the classification results of the mental health issues and the other issues. As a result, we achieved an F1-score of 0.646. The best performance was obtained using only clients' messages for training and testing and using TF-IDF as features. Even when clients' and counselors' messages, or only counselor' messages, were used for training and testing, the best scores were obtained when TF-IDF was used as a feature. We also analyzed the relationship between the F1 scores and the number of conversational messages in the best feature set, TF-IDF. Fig. 1 shows the relationship between the F1 scores and the degree of the dialogue's progress. For the estimation, messages from the beginning to the middle of the dialogue were used. The setting using the client's messages showed high scores earlier than the counselor's one, and it reached equilibrium in about ten messages.

B. Important Words

We analyzed with TF-IDF the most important words as features. The average of TF-IDF weights in all documents on mental health issues was calculated for each word. Table IV shows the ten most important words. A pseudo-dialogue example is shown below, based on the dataset. The parts corresponding to the words in Table IV are shown in bold.

Client: *"The current environment is too **tough**. I can't go **out** because of the coronavirus, and I'm constantly forced to endure it."*

Counselor: *"I see. I'm sorry that you are having such a **tough** time."*

Client: *" I'm not **watching** much TV **news** these days. Because all they talk about is the spread of the **infection** and the **increasing** number of deaths, which makes me anxious."*

Counselor: *"Although we need such factual information, I also understand that you don't want to see depressing **news** that simply fuels your **fears**."*

Client: *"Thank you very much. It is **calming** to know that this is a place where I can be heard."*

The counselor tends to repeat the words stated by the client, indicating that the counselor is aware of the important words articulated by the client. This tendency implies

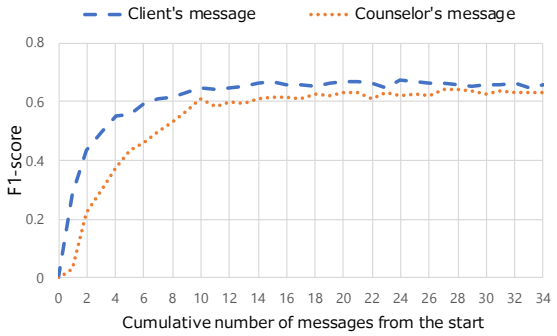

Fig. 1. Relationship between number of dialogue messages used for training and F1 scores: Degree of dialogue progress is shown on x-axis.

the importance of considering both the counselor and the client's messages and focusing on the words repeated by the counselor to identify the important words.

Some words in Table IV indicate that TF-IDF can leverage the clients' mental state for classification. "Palpitations", "Scary," "Though," "Calm," and "Fear" can be perspective as words that suggest a mental state. On the other hand, it is difficult to understand the relationship between mental health issues and words containing "News," "Television," and "Infection." The data collection situation under the COVID-19 pandemic may have affected these results, as in the pseudo-dialogue example. We are interested in exploring the relationship between mental health issues and the social situation for future work.

IV. CONCLUSION

We found that the detection model using TF-IDF as the feature and client-side messages as data is higher performance and interpretability. The next step is to implement a system that assists counselors. From the classifier of this study, it is possible to build a system that presents the predicted value of whether the client has a mental health issue and the words that are the evidence of the classification of the client.

We consider that this research can be applied not only to the support of counselors but also to the research of conversation agents that automatically provide mental health care, such as cognitive behavior therapy [16], [17]. Our detection model is promising to enable conversational agents to detect mental health issues. We expect that the agents detect mental health issues and adaptively care for the user.

ACKNOWLEDGMENT

This work was supported by JST CREST Grant Number JPMJCR19A5, Japan, Osaka Prefectural Government and Kansai Counseling Center.

TABLE IV

TOP 10 WORDS WITH HIGH WEIGHTING CALCULATED WITH TF-IDF: WE TRANSLATED ORIGINAL JAPANESE INTO ENGLISH.

Client	Counselor	Client and counselor
Fear	Palpitations	Scary
Scary	Scary	Look
Palpitations	Firmly	Infection
Infection	Calm	Palpitations
Television	Out	Tough
Look	When	Calm
Out	Caution	News
Tough	Doctor	Out
Many	Important	Television
Increase	Great	Increase

REFERENCES

[1] M. Dowling and D. Rickwood, "Online counseling and therapy for mental health problems: A systematic review of individual synchronous interventions using chat," *Journal of Technology in Human Services*, vol. 31, no. 1, pp. 1–21, 2013.

[2] C. Rauschenberg, A. Schick, D. Hirjak, A. Seidler, I. Paetzold, C. Apfelbacher, S. G. Riedel-Heller, and U. Reininghaus, "Evidence synthesis of digital interventions to mitigate the negative impact of the covid-19 pandemic on public mental health: rapid meta-review," *Journal of medical Internet research*, vol. 23, no. 3, p. e23365, 2021.

[3] Y. Sugihara and T. Miyata, *SNS Kaunsering Handobukku (The SNS Counseling Handbook)*. Seishin Shobo,Ltd., 2020.

[4] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[5] N. Cummins, A. Baird, and B. W. Schuller, "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Methods*, vol. 151, pp. 41–54, 2018.

[6] J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker, and M. Breakspear, "Multimodal assistive technologies for depression diagnosis and monitoring," *Journal on Multimodal User Interfaces*, vol. 7, no. 3, pp. 217–228, 2013.

[7] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou, "Multimodal and multiresolution depression detection from speech and facial landmark features," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pp. 43–50, 2016.

[8] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 7, 2013.

[9] T. Gui, L. Zhu, Q. Zhang, M. Peng, X. Zhou, K. Ding, and Z. Chen, "Cooperative multimodal approach to depression detection in twitter," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 110–117, 2019.

[10] T. Althoff, K. Clark, and J. Leskovec, "Large-scale analysis of counseling conversations: An application of natural language processing to mental health," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 463–476, 2016.

[11] Y. Xu, C. S. Chan, C. Tsang, F. Cheung, E. Chan, J. Fung, J. Chow, L. He, Z. Xu, and P. S. Yip, "Detecting premature departure in online text-based counseling using logic-based pattern matching," *Internet interventions*, vol. 26, p. 100486, 2021.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[14] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," tech. rep., 2015.

[15] R. Van Hout and A. Vermeer, "Comparing measures of lexical richness," *Modelling and assessing vocabulary knowledge*, vol. 93, p. 115, 2007.

[16] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial," *JMIR mental health*, vol. 4, no. 2, p. e19, 2017.

[17] K. Shidara, H. Tanaka, H. Adachi, D. Kanayama, Y. Sakagami, T. Kudo, and S. Nakamura, "Automatic thoughts and facial expressions in cognitive restructuring with virtual agents," *Frontiers in Computer Science*, p. 13, 2022.