Classification of user sentiment and utterance in daily dialog

Kwanglo Lee[1]

Abstract— **Understanding underlying emotions and utterance is the starting point for continuing daily conversations. Current study trained two different dataset labeled with 7 sentiment and 7 utterance to classify underlying sentiment and utterance from daily dialogs.**

Machine learning models (Naive Bayes, Logistic Regression, Support Vector Machine, Random Forest) were tested with accuracy and F1-score as performance reference for deep learning models. Naive Bayes and Logistic Regression model performed best among them and Logistic Regression showed slightly better results in both sentiment (Test accuracy:0.49, F1:0.49) and utterance(Test accuracy:0.66, F1:0.54).

In deep learning models (CNN-FastText, Huggingface BERT-multilingual), Huggingface BERT performed better with test accuracy (Sentiment:0.49,Utterance:0.62). Testing classification results with new inputs was also conducted using CNN-FastText model.

Dataset extension using semi-supervised learning and testing various models along with active parameter tuning will result enhanced performance in the future.

I. INTRODUCTION

Primary research tool in psychology and psychiatry researches are quantitative methods like questionnaires. Unfortunately, qualitative research requires great amount of time and resources in its process and analysis. Still, qualitative research is essential when studying psychological or psychiatric features. Especially in counseling, although initial criteria is set according to first visit questionnaire, qualitative analysis based on counseling results holds the most critical portion.

The most important part of the results of such counseling is dialogue. However, in the existing counseling method, the counselor records and summarizes the conversation contents directly and identify the flow and intention of the conversation by hand. One of the things that counselors are interested in is the specific expression of feeling and speech intentions about certain event or topic. However, as mentioned earlier, it takes too much effort to record and analyze the whole counseling session.

As the advance of natural language processing(NLP) technology, classifying specific sentiment and intention using pre-label data became possible by many supervised learning techniques. Considering such previous results, it would be possible to automatically detect speaker's intention and emotion underlying the daily log and present the results to its stakeholders.

[1]Kwanglo Lee is M.S. candidate in Department of Human Factors Engineering, Ulsan National Institute of Science and Technology, 50, UNIST-gil, Korea, Republic of

II. RELATED WORK

A. An Empirical Comparison of Machine Learning Models for Classifying Emotions in Korean Twitter

The paper completed the emotion classification utilizing machine learning models(Support Vector Machine, Naive Bayes). In this study, sentiment analysis was conducted using Twitter which is one of the most popular microblogs in the market. With two machine learning models, three different tokenizing(morpheme, bi-gram, tri-gram) was applied. Dataset was consisted with 1,333 tweets randomly selected from total 1,563,944 tweet dataset and labeled positive, negative and neutral manually. Results indicated that SVM show better performance compared to NB in binary sentiment classification.

B. Multi-class Sentiment Classification on Twitter using an Emoji Training Heuristic

This paper used twitter emoji such as "Huggingface" to automatically label large amount of tweets. Semantic map of twitter emojis built by instagram team was adapted to convert emojis to sentiment labels. Four different sentiments(Sad, Anger, Fearful, Happy) were tested using Naive Bayes method with uni-gram and bi-gram. By reviewing the result of confusion matrix, "Sad" was commonly misclassified as "Anger" and other relationship between sentiments were also found.

C. Speech Intention Understanding in a Head-final Language: A Disambiguation Utilizing Intonation-dependency

Utterance of daily language was using both text and spoken script. Concept of intonation-dependency was introduced to overcome the limitation of interpreting utterance of text data. Main classification was divided into three categories, fragment(Under-specified intention), clear-cut(Intention) and intonation-dependent(Audio-aid required). Additionally, multi-modal analysis for audio(CNN/RNN-based) and transcript(RNN-based) was used to classify user intention. Text-only using large text corpus showed best performance in Accuracy and F1-score followed by Text(large) + Multi-modal model.

III. METHOD

A. Data collecting

1) Korean emotion-labeled singular dialog dataset: Multi-sentiment dataset was collected from AI Hub which is hosted by Korea Ministry of Science and ICT(MSIT). Dataset consist of total 38,594 daily dialogs with two columns "Sentence" for dialogs and "Emotion" for sentiment

label. Types of label were Neutral(4830), Fearful(5468), Surprise(5665), Anger(5665), Sadness(5267), Happiness(6037) and Disgust(5429). Numbers inside the parentheses are number of sentences which belongs to certain sentiment.

2) 3i4K - Intonation-aided intention identification for Korean: Utterance classification used 3i4k dataset built by SNU Spoken Language Processing laboratory (SNU SLP). Dataset consist of total 61,255 sentences with two columns "text" for each sentence and "label" for utterance label. Data was categorized into mainly three types as below and thus called as FCI dataset.

- F: Fragments (nouns, noun phrases, incomplete sentences etc.) (FRs)
- C: Clear-cut cases (statements, questions, commands, rhetorical questions, rhetorical commands) (CCs)
- I: Intonation-dependent utterances (IUs)

Specific utterance were total seven categories including Fragments/FR (6009), Statements/S (18300), Questions/Q (17869), Commands/C (12968), Rhetorical questions/RQ (1745), Rhetorical commands/RC (1087) and Intonation-dependent utterances/IU (3277). Numbers inside the parentheses are number of sentences which belongs to certain utterance.

B. Data preprocessing

All datasets dropped unnecessary columns except for those with target text and label, then went through tokenization and removed stopwords.In tokenization, Okt from KoNLPy was use. After that, dataset was separated into train, test, and valid set for future application in machine learning and deep learning models.

C. Machine learning

Word Embedding used in machine learning methods was pre-trained fastText vectors for Korean. For vectorizer, various famous vectorizers were tested which includes Count Vector, TF-IDF, Ngram, Character level TF-IDF.

Applied machine learning models were Naive Bayes(NB), Logistic Regression(LR), Support Vector Machine(SVM) and Random Forest(RF) which are widely used in other NLP classification studies. Models that show best performance in accuracy and F1-score will be compared with deep learning models.

D. Deep learning

Word embedding used in deep learning model was also pre-trained FastText vectors for Korean to show consistency in comparison. Convolution Neural Network(CNN) was tested for reference and Transformer model using hugging face multi-lingual BERT developed by Huggingface was adopted for comparison. In addition, real-time classification with evaluation model was done for further researches.

IV. Result

A. Machine learning techniques

1) Multi-sentiment: Naive Bayes, Logistic Regression

Character Level TF -IDF		
	Naive Bayes	**Logistic Regression**
Validation Accuracy	0.49	0.50
Test Accuracy	0.49	0.49
F1-score	0.48	0.49

Fig. 1. Performance of machine learning models in multi-sentiment classification

Figure 1 shows the validation, test accuracy and F1-score of machine learning methods in multi-sentiment classification. Both models show similar performance while logistic regression has slightly higher performance compared to Naive Bayes. Confusion matrix was generated to compare performance within each labels.

	Neutral	**Fearful**	**Surprise**	**Anger**	**Sadness**	**Happiness**	**Disgust**
Neutral	230	158	328	234	99	166	234
Fearful	70	918	210	50	261	71	60
Surprise	92	215	989	115	115	133	110
Anger	59	89	226	854	77	45	350
Sadness	26	284	115	69	943	81	52
Happiness	60	56	201	51	105	1287	51
Disgust	119	94	257	526	89	68	476

Fig. 2. Confusion matrix of Naive Bayes(Multi-sentiment)

	Neutral	**Fearful**	**Surprise**	**Anger**	**Sadness**	**Happiness**	**Disgust**
Neutral	376	103	266	221	78	160	245
Fearful	115	859	218	51	256	62	79
Surprise	167	170	937	108	106	129	152
Anger	128	66	183	776	63	44	440
Sadness	70	249	115	75	923	84	64
Happiness	105	44	157	52	92	1292	69
Disgust	210	72	202	452	78	68	547

Fig. 3. Confusion matrix of Logistic Regression(Multi-sentiment)

In confusion matrix, row index refers to actual labels and column index indicate predicted labels. As in figure 2 and 3, there were high similarities in confusion matrix of both model. Specifically, Neutral tend to be misclassified as Disgust, Fearful as Surprise and Sadness, Surprise as Neutral and Fearful, Anger as Neutral and Disgust, Sadness as Fearful, Happiness as Neutral and Surprise and Disgust as Neutral and Anger. These results show that there are unknown underlying correlation between word vectors of certain sentiments.

2) Utterance: Naive Bayes, Logistic Regression

Figure 4 shows the validation, test accuracy and F1-score of machine learning methods in utterance classification. Logistic regression showed noticeable difference between two models and both models performed better and compared to multi-sentiment analysis. Confusion matrix was generated to compare performance within each labels.

Figure 5, 6 show the performance difference in individual utterance between two models. Even though Logistic

Character Level TF -IDF		
	Naive Bayes	**Logistic Regression**
Validation Accuracy	0.61	0.71
Test Accuracy	0.58	0.66
F1-score	0.54	0.64

Fig. 4. Performance of machine learning models in utterance classification

Regression performed better in overall, there shows lack of data in RQ, RC and IU. In addition, Statement tend to be misclassified as Question and Command, Question as Command and Command as Question. Similar to sentiment analysis, there seemed to be unknown correlation within utterances that needs to be took into account.

	FR	S	Q	C	RQ	RC	IU
FR	21	485	66	26	0	1	1
S	3	1518	210	90	1	1	7
Q	0	244	1285	253	2	0	2
C	1	324	352	618	0	1	0
RQ	1	130	23	3	13	0	4
RC	0	70	14	6	0	18	0
IU	0	160	41	21	0	1	104

Fig. 5. Confusion matrix of Naive Bayes(Utterance)

	FR	S	Q	C	RQ	RC	IU
FR	553	54	11	7	0	0	1
S	83	1450	176	103	4	3	11
Q	32	257	1266	225	4	0	2
C	23	297	359	611	1	3	2
RQ	7	109	19	12	21	0	6
RC	4	55	8	7	0	33	1
IU	17	127	38	22	0	1	122

Fig. 6. Confusion matrix of Logistic Regression(Utterance)

B. Deep learning techniques

1) Convolution Neural Network: Multi-modal Model

CNN - FastText Embedding		
	Multi-sentiment	**Utterance**
Validation Accuracy	0.42	0.66
Test Accuracy	0.41	0.43
F1-score	1.00	1.00

Fig. 7. Performance of CNN in multi-modal analysis

Figure 7 shows the result of CNN - FastText model for both tasks. Similar to machine learning results, utterance showed higher performance compared to multi-sentiment classification.

Figure 8, 9 are confusion matrix of CNN-FastText model. Strong similarity between sentiment is also found in figure 8 which is similar results to machine learning models. Still, there are much more mis-classified cases compared to machine learning methods.

Figure 9 reveals that there are high mis-classification in major labels(FR, S, Q) in Utterance. In addition, RQ, RC and IU hold very little portion of total dataset.

	Neutral	Fearful	Surprise	Anger	Sadness	Happiness	Disgust
Neutral	646	141	182	306	161	104	229
Fearful	159	1136	77	115	74	82	168
Surprise	173	34	729	156	363	57	188
Anger	202	47	104	899	82	176	130
Sadness	195	65	507	141	441	44	236
Happiness	112	76	106	371	67	740	108
Disgust	230	135	221	230	218	78	337

Fig. 8. Confusion matrix of Multi-sentiment(CNN)

	FR	S	Q	C	RQ	RC	IU
FR	3998	709	460	264	74	45	30
S	616	3759	819	40	44	39	5
Q	821	900	2047	56	13	14	8
C	182	39	33	1506	10	4	6
RQ	324	130	76	32	397	6	3
RC	274	78	27	25	24	101	1
IU	165	31	24	13	3	3	99

Fig. 9. Confusion matrix of Utterance(CNN)

2) Huggingface BERT: Separate comparison

Huggingface BERT		
	Multi-sentiment	**Utterance**
Validation Accuracy	0.51	0.73
Test Accuracy	0.49	0.62

Fig. 10. Performance of Huggingface BERT

Figure 10 is result of Huggingface BERT multi-lingual model. Unlike CNN-FastText, it was tested separately due to technical issues. Similar to previous performance tables, Utterance performed better and Multi-sentiment analysis.

C. New user inputs

Testing new inputs were done only in CNN model. Figure 11 shows the results of new input testing. Even though overall performance of the model was not remarkable, it correctly classified newly suggested dialog showing potential use in future study.

V. Discussion

The highlight of this study is that it implemented classification model that detects the latent sentiment and utterance of everyday conversation by combining the multi-sentiment classification and utterance recognition. Moreover, current model successfully classified sentiment and utterance from newly inputted sentence.

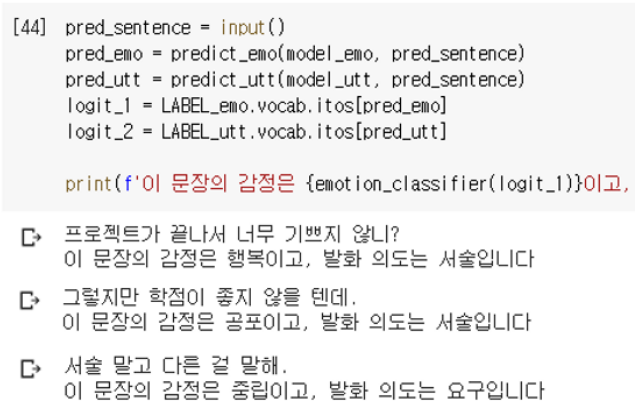

Fig. 11. New input example in multi-modal analysis using CNN

If these results are improved to the level that can be applied to actual counseling, many of the current manual tasks will be converted to computer-aided tasks, which will enable more effective mental healthcare service. However, despite these achievements, this study has several limitations.

A. Limitation

The first limitation is lack of data size. In this study, logistic regression was better than other deep learning models. In this regard, a larger dataset is needed to maximize the advantages of the deep learning algorithm. To solve this problem, sub-goal of the study was to expand the dataset through semi-supervised learning, but it was not implemented due to technical issues.

In addition to dataset size problems, imbalance between labels is another factor that decreased the performance. RQ, RC, and IU have overwhelmingly fewer numbers than other labels in the utterance dataset.Although it is difficult to secure data balance since such utterance have rare frequency in actual conversations, it is necessary to balance the label to until certain level for sufficient accuracy.

The next limitation is the absence of multi-modal analysis using various models. In this study, CNN-FastText and Huggingface BERT Multilingual models were used. Both of these were selected because they are more suitable than many other popular models in Korean NLP tasks. Still, they are not enough compared to state-of-art model like K-BERT by ETRI.

Therefore, testing various model and embedding vectors will help in achieving enhanced performance in the future. Also, due to the time limit, various parameter tunings could not be done. This is also thought to be a problem that can be improved through repeated experiments.

B. Further study

For further study, it is necessary to expand the dataset through adapting methods like semi-supervised learning as mentioned above. In addition, this study utilized the same CNN+FastText model for both sentiment and utterance dataset. If more suitable multi-modal classification model can be used for each task, higher performance is expected. Finally, it is necessary to secure classification accuracy and establish classification criteria for new input data. By doing so, it will boost not only identifying the user utterance more effectively, but also help to expand the data set.

REFERENCES

[1] Lim, J. S., Kim, J. M. (2014). An empirical comparison of machine learning models for classifying emotions in korean twitter. Journal of Korea Multimedia Society, 17(2), 232-239.

[2] Hallsmar, F., Palm, J. (2016). Multi-class sentiment classification on twitter using an emoji training heuristic.

[3] Cho, W. I., Lee, H. S., Yoon, J. W., Kim, S. M., Kim, N. S. (2018). Speech intention understanding in a head-final language: A disambiguation utilizing intonation-dependency. arXiv preprint arXiv:1811.04231.

[4] Park, S., Byun, J., Baek, S., Cho, Y., Oh, A. (2018, July). Subword-level word vector representations for Korean. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2429-2438).

[5] Cho, W. I., Cheon, S. J., Kang, W. H., Kim, J. W., Kim, N. S. (2018). Real-time automatic word segmentation for user-generated text. arXiv preprint arXiv:1810.13113.

[6] Zhou, M., Duan, N., Liu, S., Shum, H. Y. (2020). Progress in Neural NLP: Modeling, Learning, and Reasoning. Engineering.