Comparing Performance of Pre-trained Bidirectional Language Models

Hyeon Jeong Byeon
Ewha Womans University
cat@ewhain.net

Uran Oh*
Ewha Womans University
uran.oh@ewha.ac.kr

Abstract

Patients often share information about their symptoms online by posting on web communities and SNS. While these posting data have been proven to be useful for improving psychological therapy experiences, little has known if and how the same approach can be applied to Korean. This paper investigates the performance of bidirectional language models. Results show that both multi-lingual BERT model and KoBERT (Korean BERT) model perform well on binary sentiment classification, reaching an accuracy of 90%. In addition, bcLSTM models outperformed on emotion recognition that classifies casual texts into Paul Ekman's six emotions than positive/neutral/negative sentiment analysis. Through this research, we concluded that in order to utilize sentiment analysis models in psychological therapies, additional layer that detects certain psychological symptoms are necessary. As our future task, we are looking forward to propose a new deep learning model that detects emotion disorders.

Keyword

Sentiment Analysis, Emotion Classification, Natural Language Processing, Language Models

1. Introduction

According to a national survey conducted by the Pew Research Center's Internet & American Life Project, almost 80% of Internet users, which is about 93 million distinct individuals in the United States, surf the internet to gain medical information [1]. Many of them go so far as to diagnose themselves or share information about their symptoms online, creating a class of records called ePAT(electronic Patient-Authorized Text). Prior researches show that analysis of ePAT can be an effective means for identifying non-pharmaceutical forms of intervention [2, 3]. For instance, Shing et al. [3] investigated online postings using data from Reddit (reddit.com), a popular website for anonymous discussions on a wide variety of topics, particularly a subreddit called "SuicideWatch." The study suggests that these online discussion forums can be an effective source to gather ePAT datasets for machine learning applications, when natural language processing techniques such as topic model posteriors [25], word embeddings, and emotion features [26] etc. are used to engineer the features of the dataset. Other study has explored the use of natural language processing (NLP) techniques for improving the delivery of mental health services [5]. We believe that this type of analysis would be particularly useful for South Korea, since mental health services are difficult to access and heavily stigmatized when compared to North American and European countries [21]. Inspired by the fact that little has known about emotion and stress detection in Korean text, we examined if and how Korean online postings can play an important role as a valuable source for several services, such as deep learning models that detects mental health distressers and symptoms, and empathy-driven conversational artificial intelligence.

To explore the feasibility of this application, we investigated four sentiment analysis approaches based on deep learning models; LSTM model [9], bcLSTM model[10], BERT model [11], and KoBERT model [12]; and compare their performance on a range of relevant text datasets; IMDb review datasets [8], Naver sentiment movie corpus v1.0 [13], Reddit stress data [14], and the TV series <Friends> data [15].

*교신저자

Results showed that both BERT models and LSTM models well-performed on positive negative sentiment analysis by reaching an accuracy of 89%. KoBERT model as well showed high performance on sentiment analysis with an accuracy of 90%. However, the KoBERT model poorly performed under conditions when the source of the training dataset and the test dataset was given different. From the results of two conditions in the experiment (i.e., training the KoBERT model with stress data and testing it with positive/negative data and training the KoBERT model with positive/negative data and testing it with stress data), we concluded that the presence of one's stress has low correlation - or completely independent - with the patients' positive negative language usage. All things considered, we concluded that binary sentimental analysis deep learning models are inapplicable for psychological therapies. As a future work, we plan to improve the performance of deep learning models for psychological usage by inserting additional layers that detect certain psychological symptoms such as depression and bipolar disorder besides one's emotions.

2. Related Works

2.1 Identifying Emotions from User-generated Texts

According to Derks et al. [23], most of the social media platforms encourage people to express their emotions, and constantly update their experiences, feelings and thoughts to other users. Due to the growth of textual messaging platforms such as WhatsApp[1] and Twitter[2], there are a countless amount of text data to use as a source for natural language processing. Chatterjee et al. [7] proposed a deep learning based approach called "Sentiment and Semantic Based Emotion Detector (SS-BED)" to detect emotions in textual dialogues at Twitter. Also, Shivhare et al. [22] discussed several methods in emotion detection from text; including keyword spotting technique, lexical affinity method, learning-based methods, and hybrid methods. Meanwhile, Thelwall et al. [24] proposed an emotion detection algorithm SentiStrength to extract emotions from casual text using data from MySpace.[3]

This paper conducted a comparison study of sentimental analysis in order to determine the applicability of former sentimental analysis algorithms on ePAT data, unlike many other papers that are focused on detecting the emotions of randomly gathered text.

2.2 Deep Learning Models for Emotion Detection

Various deep learning models had been used for detecting emotions. Proven by State-of-The-Art deep learning algorithms, LSTM models and BERT models show good performance [19].

LSTM Models [9] were proposed to solve the vanishing Gradient problem of RNN (Recurrent Neural Networks) models [20] whose outputs are highly dependent on the previous calculations. LSTMs loop through a row of data, persevering and clustering a class of the working memory over multiple times. Among diverse LSTM Models, bcLSTM(bidirectional contextual LSTM), proposed by Poria et al. [10] is notable for using contextual information as well as a bi-directional RNN model. Meanwhile, Devlin et al. [11] introduced a new language model named BERT, which stands for Bidirectional Encoder Representations from Transformers. The biggest difference between BERT and other language models is that it is designed to pre-train deep bidirectional representations from unlabeled text by mutually conditioning on both sides of the context in all layers. This allows BERT to create a wide range of models and solve a wide range of tasks.

In this paper, we summarized distinctive characteristics of four models in particular: LSTM (Long Short Term Memory Networks), bcLSTM(bidirectional contextual LSTM), BERT (Bidirectional Encoder Representations from Transformers), and KoBERT (BERT for Korean), which is designed to improve the accuracy of multi-lingual BERT models on Korean language corpora [12].

[1] https://web.whatsapp.com/

[2] https://www.twitter.com

[3] https://myspace.com/

3. Experiments

We conducted this experiment in order to determine a deep learning model that shows the highest accuracy on sentiment analysis for Korean. Furthermore, we tested the applicability of existing sentiment analysis models on stress data.

3.1 Datasets

- 3.1.1 Naver Sentiment Movie Corpus

“Naver sentiment movie corpus v1.0” [13] is a Korean-language dataset consisting of reviews scraped from “Naver Movies,” a popular South Korean online venue for anonymous reviews and ratings on movies distributed in the South Korean market. The dataset was constructed based on the method noted in Maas et al. [8]

Each file consists of three columns: id, document, and label. The “id” column contains the unique number ID assigned to the reviewer within Naver’s databases. The “document” column contains the text from the actual review, which is typically limited to 140 characters. The “label” column contains the sentiment label for the review, which has been rated on a Likert scale from 0 (most negative) to 10 (most positive). Reviews labeled 5 through 8 were considered to be “neutral” reviews and were excluded from this dataset, which leaves reviews with 0-4 as “negative” and 9-10 as “positive”. As a result of that, the dataset was sampled equally on each sentiment class. Among 200K reviews in total, we randomly assigned 50K for testing, and the remaining 150K reviews for training.

- 3.1.2 Reddit Stress Data

Turcan et al. [14] introduced “Dreaddit”, a text corpus of lengthy social media data for the identification of stress. The dataset consists of 190K posts from five different Reddit communities (i.e., r/abuse, r/anxiety, r/financial, r/PTSD, r/social) and additional 3K posts from Amazon Mechanical Turk. Since Reddit is a social media website where users post in topic-specific communities, most of the posts are long enough to detect the nuances of phenomena like stress. The authors chose five categories of subreddits which are most likely to be associated with stress; r/abuse, r/anxiety, r/financial, r/PTSD, and r/social [14]. Unlike other datasets, the Dreaddit corpus is noteworthy because it has the potential to spur development of sophisticated models of psychological stress.

- 3.1.3 TV series <Friends> data

Poria et al. [15] proposed the Multimodal EmotionLines Dataset (MELD), an extension and improvement of EmotionLines. MELD dataset is made out of dialogues from the TV series Friends. The authors extracted 13K utterances from 1,433 conversations from the show. Each utterance of the dialogue is annotated with sentiment and emotion labels, and consists of audio, visual, and textual modalities. See Section 4.3 for how the data is labeled for our experiment.

3.2 Procedure

We first compared the performance of LSTM and BERT models with IMDb review datasets [8]. Then we conducted another experiment on only BERT models, which outperformed LSTM. In particular, we compared the accuracy of the KoBERT model in various situations.

Experiments were performed in four conditions: (1) training and testing KoBERT model with Naver sentiment movie corpus, (2) training KoBERT model with Naver sentiment movie corpus and testing it with Dreaddit dataset, (3) training KoBERT model with Dreaddit dataset and testing it with Naver sentiment movie corpus, and (4) training and testing KoBERT model with Dreaddit dataset. As for Dreaddit dataset, to be trained or tested with KoBERT model, we used Google Translate[4] to create a Korean version. Note that the nuance of sentences may have been lost during the translation process.

In addition, we trained and tested the bcLSTM model using the TV series Friends dataset[15] only using the text modality, to see how the bcLSTM model works with casual texts. We used an open source project on Github called MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversation [16] to conduct the experiment.

[4] https://translate.google.com/

4. Results

4.1 LSTM Versus BERT

We first compared the performance of LSTM and BERT for sentiment analysis, trained and tested with the IMDb review dataset [8], and the accuracy results are shown in Table 1. We used open source projects through Google Colab for LSTM [17] and [18] for BERT.

Table 1. Accuracy results of LSTM and BERT models

Epoch #	LSTM	BERT
1	77.72%	88.16%
2	84.94%	89.12%
3	87.82%	89.61%
4	89.00%	89.75%
5	89.02%	89.70%

Results of both models show that increased epoch (iteration over the entire training data) leads to an increase in the accuracy of the sentiment analysis. Also, we can see that the accuracy of BERT model was 0.68% higher than that of LSTM model.

4.2 Comparing accuracy of KoBERT model with Naver Sentiment Movie Corpus and Reddit Stress dataset

We performed our experiments in four conditions by using Naver Sentiment Movie Corpus[13] and Reddit Stress dataset[14].

Table 2. Accuracy results of BERT and KoBERT models. Note that Reddit data were translated from English to Korean for the experiment.

Model	Train data	Test data	Accuracy
KoBERT	Naver	Naver	90.00%
KoBERT	Naver	Reddit	50.83%
KoBERT	Reddit	Naver	50.35%
KoBERT	Reddit	Reddit	70.80%

Results shown above on Table 2 signifies the performance of the KoBERT model in several situations.

4.3 Training and testing bcLSTM model using TV series <Friends> text data

As you can see from the dataset proposed by Shing et al. [3], ePAT datasets extracted from online discussions are casual texts. We examined two types of classification accuracy using a pre-trained bcLSTM model to compare the performance of sentiment classification and emotion classification. As Pennebaker et al. [27] have mentioned, patients' emotional words play an important role during psychotherapies. Therefore, detection of positive negative languages and emotional languages are necessary for our research. We chose TV series <Friends> text data[15] to conduct our experiments.

The first type was testing the sentiment classification; label 0 as Neutral, 1 as Positive, and 2 as Negative. The second one was testing the emotion classification of this model, based on Ekman's six universal emotions (Joy, Sadness, Fear, Anger, Surprise, and Disgust). bcLSTM model outperformed on emotion classification than sentiment classification. See Table 3 for the results.

Table 3. Accuracy results of bcLSTM

Test version	Accuracy
Sentiment Classification	48.12%
Emotion Classification	60.5%

Our experiments showed poorer performance with emotion classification than sentiment classification where the former has more number of classes. This could be due to the skewed dataset for the sentiment classification dataset where the amount of "neutral" data is twice more than "positive" or "negative" data[15].

5. Conclusion

We conducted multiple experiments with four different deep learning models (i.e., LSTM, bcLSTM, BERT, and KoBERT) with various types of data to investigate the feasibility of assessing the data for improving psychological therapy experiences, especially ePAT data.

Results in Section 4.1 showed that the accuracy of BERT model was 0.68% higher than that of LSTM model, when both models were trained and tested with an identical dataset. According to the results from Section 4.2, KoBERT model outperformed when the source of the training dataset and the test dataset were the same. This means that the presence of one's stress has low correlation with positive negative language usage. From this result, we concluded that binary sentiment anlaysis algorithms are inapplicable for detecting negative nuances of ePAT data. However, we confirmed that the bcLSTM language models are appropriate for emotion recognition which is also an important feature required in psychological therapies.

Therefore, as our future tasks, we plan to propose a new deep learning model by inserting additional layers that detect probabilities of certain emotion disorders from Korean online postings, considering unique Korean internet cultures [6].

Acknowledgement

This research was supported by UROP (Undergraduate Research Opportunities Program) Program at Ewha Womans University.

Reference

1.Fox, C., and Duggan, M. Health Online 2013. Pew Research Center Internet & American Life Project (2013), 1-55.

2.Dreisbach, C., Koleck, T. A., Boume, P. E., and Bakken, S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authorized text data. International Journal of Medical Informatics (2019), 125, 37-46.

3.Shing, H. C., Nair, S., Zirikly, A., Friedenberg M., Daumé III, H., and Resnik, P. 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. Association for Computational Linguistics (2018), 25-36.

4.Pace, B., Tanana, M., Xiao, B., Dembe, A., Soma, C., Steyvers, M., and Imel, Z.E. What About the Words? Natural Language Processing in Psychotherapy. Psychotherapy Bulletin 51, 1(2016), 17-18.

5.Wang, X., Zhang, C., Ji, Y., Sun, L., Wu, L., and Bao, Z. A depression detection model based on sentiment analysis in micro-blog social network. Pacific-Asia Conference on Knowledge Discovery and Data Mining (2013), 201-213.

6.대한민국 여론 움직이는 6대 온라인 커뮤니티. http://weekly.chosun.com/client/news/viw.asp?ctcd=C01&nNewsNumb=002580100001

7.Chatterjee, A., Gupta, U., Chinnakotla, M. K., Srikanth, R., Galley, M., and Agrawal, P. Understanding emotions in text using deep learning and big data. Computers in Human Behavior, 93 (2019), 309-317.

8.Mass, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis.
In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (2011), 142-150.

9. Hochreiter, S. and Schmidhuber, J. Long short-term memory. Neural computation, 9, 8 (1997), 1735-1780.

10. Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., and Morency, L. P. Context-dependent sentiment analysis in user-generated videos. In Proceedings of the 55th annual meeting of the association for computational linguistics, 1(2017), 873-883.

11. Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXive: 1810.04805 (2018).

12. SKTBrain/KoBERT: Korean BERT pre-trained cased (KoBERT). https://github.com/SKTBrain/KoBERT

13. e9t/nsmc : Naver sentiment movie corpus. https://github.com/e9t/nsmc/

14. Turcan, E., and McKeown, K. Dreaddit: A Reddit dataset for stress analysis in social media. arXiv preprint arXiv: 1911.00133 (2019).

15. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. Meld: A multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508 (2018).

16. Declare-lab/ MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversation. https://github.com/declare-lab/MELD/

17. Explaining Recurrent Neural Networks. https://www.bouvet.no/bouvet-deler/explaining-recurrent-neural-networks

18. Ktrain: A Lightweight Wrapper for Keras to Help Train Neural Networks. https://towardsdatascience.com/ktrain-a-lightweight-wrapper-for-keras-to-help-train-neural-networks-82851ba889c

19. Browse SoTA: Natural Language Processing: Sentiment Analysis. https://paperswithcode.com/task/sentiment-analysis

20. Recurrent Neural Networks Tutorial. http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

21. 조다혜. 한국과 미국 TV 드라마에 재현된 커뮤니케이션 양상 비교 연구. 국내석사학위논문 한국외국어대학교 대학원 (2019).

22. Shivhare, S. N., and Khethawat, S. Emotion detection from text. arXiv preprint arXiv: 1205.4944 (2012).

23. Derks, D., Fischer, A. H., and Bos, A. E. The role of emotion in computer-mediated communication: A review. Computers in human behavior, 24, 3 (2008), 766-785.

24. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. Sentiment strength detection in short informal text. Journal of the American society for information science and technology, 61, 12(2010), 2544-2558.

25. Blei, D. M., Ng, A. Y., and Jordan, M.I. Latent Dirichlet allocation. Journal of machine Learning research (2003), 993-1022.

26. Mohammad, S. M., and Turney, P. D. Crowdsourcing a word-emotion association lexicon. Computational Intelligence, 29, 3(2013), 436-465.

27. Pennebaker, J. W., Mehl, M. R., and Niederhoffer, K. G. Psychological aspects of natural language use: Our words, our selves. Annual review of psychology, 54, 1(2003), 547-577.