논문 2021-16-10

US 항공사서비스에 대한 트위터 데이터의 부정적 이유 분류
(Negative Reasons Classification for Twitter Data on US Airline Services)

국 정 우*
(Jeong-Woo Guk)

Abstract : Many companies try to analyze and utilize feedback on services. This can be used for improving service quality or marketing. Until now, most natural language processing studies have attempted to analyze emotions divided into positive, negative and neutral. However, in this work, specific negative reasons are extracted and classified. The dataset is a standard dataset from kaggle that uses tweet data for U.S. airline services. Tweets categorized as negative are labeled with 10 categories of negative reasons. The dataset was divided into train, validation, and test 8:1:1. The learning and classification process was largely divided into two stages. The first is to convert words and sentences into vector values. It is compared and analyzed using Doc2Vec and BERT (Bidirectional Encoder Representations from Transformers) models for embedding and vectorization. The second is to learn and classify sentences transformed into vectors by matching them with 10 negative reason classes. During this learning process, I converted the negative reason into a sentence and attached it to the back of the original text and made new data.I then used BERT's Next Sentence Prediction technique to allow further learning to be performed. This method was able to improve classification accuracy. For each dataset and classification method, metrics were computed, visualized, and compared.

Keywords : Topic Classification, Deep Learning, Natural Language Processing (NLP), Embedding, BERT

Ⅰ. Introduction

Major U.S. airlines are focusing on quality of service to gain an edge in competition. One of the representative ways to improve service quality is to obtain customer feedback. Therefore, many companies spend a lot of time and money trying to get accurate customer feedback. This is because companies can leverage customer feedback to improve service quality and use it for marketing. With the development of big data and machine learning technologies, there is a growing demand for solving tasks that people used to analyze and classify feedback directly with machine learning techniques. In the meantime, natural language processing for most of the feedback has focused on sentiment analysis, which is classified as positive and negative. For example, a study has been conducted to compare and analyze machine learning techniques that classify them as positive, negative, and neutral based on feedback data for U.S. airlines collected via Twitter [1].

However, airlines and other companies are increasingly demanding more specific analysis results. Companies want to go one step further from classifying text as positive and negative and extract the cause. By extracting reasons, cause analysis is possible and faster response and utilization is expected. In this paper, techniques for classifying into 10 negative reasons will be compared based on feedback data on US airlines collected from Twitter. I used the Doc2Vec [2], BERT (Bidirectional Encoder Representations from Transformers) model [3] for vector transformation, and classification techniques used Gradient Boosting [4], BiLSTM (Bidirectional Long-Short Term Memory) [5], and BERT. These models were used to learn and classify, and result metrics, including accuracy for classification, were visualized. The key idea is to make a negative reason into a sentence and attach it to the back of the original sentence and learn the inter-sentence relationship in the BERT learning process. These methods are simple but powerful. Negative reason classification showed 0.63 (Accuracy) and 0.63 (F1 score) in the BERT model. However, it was difficult to verify accurate experimental results because there was not enough data for class-specific learning. Therefore, similar negative reasons were integrated to balance data and

*Corresponding Author (jw.kook@yonsei.ac.kr)

Received: Feb. 13, 2021, Revised: Mar. 12, 2021, Accepted: Mar. 31, 2021.

Jeongwoo Guk: Yonsei University (M.S.)

© IEMEK J. Embed. Sys. Appl. 2021 Apr. 16(2) 73-77
ISSN : 1975-5066
http://dx.doi.org/10.14372/IEMEK.2021.16.2.73

conduct experiments. As a result, the prediction indicators showed 0.69 (Accuracy) and 0.68 (F1 score) in the BERT model.

II. Related Work

The work of this paper is inspired by previous works on natural language processing. Feedback datasets for U.S. airlines collected on Twitter were classified and analyzed as positive, negative, and neutral using machine learning techniques [1]. Previous work transforms text into vectors based on Doc2Vec model [2]. The dataset is compared and analyzed using a total of seven classification techniques. (Decision Tree Classifier, Random Forest Classifier and many other methods.) Accuracy indicators by classification model were displayed and the proportion of positive, negative and neutral by airline was visualized. BERT [3] is a deep learning model with excellent natural language processing performance. This model is extended in Transformer architecture [6]. Using a large amount of corpus, pre-training models learn words and sentences in both directions. BERT has shown high performance in classification and word prediction in natural language processing tasks. A detailed description of the models used can be found in the paragraph below.

1. Doc2Vec

Doc2Vec shows high performance when embedding sentences and documents from Word2Vec to an extended model [2, 7]. The Doc2Vec model can be divided into two main types. The vector model combining word vectors and paragraph matrices is PV-DM (Distributed Memory Version of Paragraph Vector). Except for word vectors, the model using only the paragraph vector is PV-DBOW (Distributed Bag Of Words version of Paragraph Vector). A paragraph vector is a vector transformation of a sentence or document, which is easy to express not only the relationship between words but also the relationship between sentences. I performed embedding using PV-DM model.

2. GradientBoosting

GradientBoosting [4] is one of the Ensemble models that combine multiple Decision Trees to find the optimal classification. As the learning progresses, gradients are derived in the direction of reducing the loss function and combine them to generate the effect of the ensemble model. To find the point where Loss Function is minimized, the algorithm differentiates loss function as a model function learned to date. Xgboost [8] is a library implemented to enable Gradient Boosting algorithms to run in distributed environments. It is widely used as a classification technique, showing efficient learning rates even in large-scale data. I used Xgboost to classify sentences into multiple classes.

3. BiLSTM

LSTM [9] is a model that introduces cell state to solve Gradient problems. In the learning process, Gradient is efficiently applied to learning progress by combining information from the previous point of time and information from the present point of time. However, if the sentence length is long and the layer is deeper, the loss of information increases, resulting in 'Bottle-Neck Problem'. To address this, BiLSTM [5] applies the previously introduced Attention [6], which draws attention to the most meaningful inputs in prediction. Furthermore, a bidirectional network that considers both forward and backward directions is applied to improve learning performance. I used the BiLSTM model for learning and classification.

4. BERT

BERT performs Dynamic Embedding with a model based on Transformers [3, 6]. Dynamic Embedding means to be given different embeddings depending on the meaning of the word. Even the same words are converted to different vectors depending on the position or meaning in the sentence. Transformers is a natural language processing model based on Attention, as indicated by the title 'Attention Is All You Need' of the paper. Attention is a method of expressing the most affected words within encoders and decoders. BERT is designed to extend part of Transformers to give a self-attention effect.

These methods allow us to have different embedding values depending on the position in the sentence, even if they are the same words. This is the biggest difference from Word2Vec-based models and is a good way to resolve redundancy. BERT pretrain large amounts of corpus based on the preceding method. Then perform sentence and word embeddings of datasets based on pretrained models.

BERT is divided into pre-training and fine-tuning phases. During the pre-training phase, model learn a large amount of corpus consisting of Wikipedia and Book Corpus. Then embedded the tweet text based on the pretrained model in the fine-tuning phase and learn and predict the class. One of the features that BERT differs from other models is that it is Deeply Bidirectional [3].

BERT adds an [SEP] identifier that represents the end of the sentence and a [CLS] identifier that represents the class before the sentence during the Embedded process. Then use two methods to enable Bidirectional learning. The first is the Masked Language Model, which performs bidirectional learning by masking words from random locations in sentences and proceeding with learning to predict them. The second is Next Sentence Prediction, which divides the sentence into two sentences based on the [SEP] token and predicts the next sentence. In this paper, I perform learning and classification by focusing on the learning layer that predicts the next sentences based on BERT.

III. Dataset

1. Dataset Preparation

The dataset used in this study was taken from Standard Kaggle Dataset : Twitter US Airline Sentiment released by CrowdFlower. According to Kaggle's explanation, volunteers collected feedback tweets about the six major U.S. airlines and classified positive, negative and neutral based on them. Volunteers were asked to categorize the causes of negative tweets into 11 categories. The Table 1 shows the number of tweets per negative reason. The labels are divided into 10 categories and indicate the reasons for the negative experience in airline services in sentences. The dataset of this task also included positive or neutral tweet data. Also labeled tweets about positivity and neutrality as 'Negative reason is none.' In this paper, datasets are divided into 8:1:1 ratios for training, validation, and testing.

2. Dataset Preprocessing

Data preprocessing is essential to obtain more accurate analysis results. It is necessary to eliminate noise that interferes with analysis and prediction and to refine only the necessary parts. pre-processing allows refined data to produce more accurate and reliable analysis results. Text that can act as noise in learning negative reason classification is replaced or removed. In the tweet text, the site address included for the link was replaced with 'URL' and the phone number with 'TEL'. The abbreviated expression was replaced by the original sentence. Table 2 shows abbreviation and alternative sentences. In this paper, learning is conducted based on English and classification is predicted. Therefore, non-English words were removed. The BERT model learns sentence relationships by dividing the two sentences into [SEP] tokens in the pre-training phase. Sentence relationship learning is possible even at the fine-tuning phase.Each tweet text has a delimiter added between the two sentences for better learning effects.

Table 1. Negative Reason Distribution of Tweets

Negative Reason	Tweet Count
None (Positive or Neutral)	5462
Customer Service Issue	2910
Late Flight	1665
Can't Tell	1190
Cancelled Flight	847
Lost Luggage	724
Bad Flight	580
Flight Booking Problems	529
Flight Attendant Complaints	481
Longlines	178
Damaged Luggage	74

Table 2. Alternative Sentences for Abbreviated Expressions

Abbreviated Expression	Alternative Sentence
n't	not
‘re	are
‘s	is
‘d	would
‘ll	will
‘t	not
‘ve	have
‘m	am
won't	will not
can't	can not

Ⅳ. Method

1. Addtional Sentences

I performed fine-tuning phase based on BERT-BASE-CASED of BERT, a pre-trained model for English. All parameters applied the standard parameters tested in the [3]. Negative reason can be converted to sentences respectively. Within the fine-tuning phase, negative reasons were converted and added to the sentence after the original sentence. For example, in Fig. 1, the original tweet text is followed by the sentence '[SEP] Negative Reason is Bad Flight'. The generated data is added to the existing dataset. This addition of the

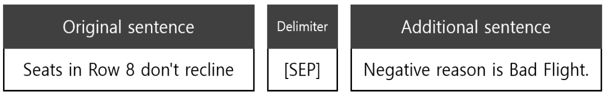


Fig. 1. Additional Sentence

sentence produces one more data in addition to the original sentence. As a result, the number of data doubled. Therefore, the BERT model further learns sentences containing negative reasons.

2. Next Sentence Prediction

Within the fine-tuning phase, sentences after the [SEP] token were replaced with other sentences with a 50% chance. Then predicts the sentences behind and proceeds with learning in a direction that minimizes loss. This is the same as learning to predict sentences made up of negative reasons added earlier. For example, in Fig. 1, the Additional Sentence part is changed to 'Negative reason is Late Flight'. Then return True if it is the same as the predicted sentence, False if it is different, and compare it to the correct answer. Learning proceeds with weight updates in a way that minimizes losses. This process is directly related to matching the classes that we want to classify. Therefore, it is an important process for increasing accuracy.

V. Experiment and Evaluation

Of the total 14640 tweets, 9178 were classified negatively, and the reasons for this were written. Positive and Neutral tweets were labeled negative reason as 'Negative reason is None'. The dataset was experimented with the Embedding Model and Classification Methods presented earlier. Models have been repeatedly trained up to 10 times. Table 3 is a performance evaluation indicator for classification models. And Table 4 shows model-specific classification result figures. Precision (1) is the number divided by the number classified by the model as True by the number of correct answers that are True. Recall (2) is the number that the model classifies as True divided by the number that the correct answer is True. F1-score (3) is the harmonic mean of Precision and Recall, and was used because the data between classes were imbalanced. F1 score allows us to verify the performance of the model on unbalanced data. If different models were used for Embedded and Classification, the + symbol was used. The result of not adding a negative reason converted to a sentence was denoted as Standard.

Table 3. Confusion Matrix

	Actual True	Actual False
Predicted True	True Positive	False Positive
Predicted False	False Negative	True Negative

Table 4. Classifier Performance Indicators

Embedding & Classifier	Accuracy	Precision	Recall	F1 Score
Doc2Vec+GradientBoost	0.411	0.411	0.411	0.411
Doc2Vec+BiLSTM	0.374	0.373	0.373	0.373
BERT+GradientBoost	0.437	0.437	0.437	0.437
BERT+BiLSTM	0.535	0.582	0.499	0.536
BERT (Standard)	0.620	0.620	0.620	0.626
BERT (Additional Sentences)	**0.636**	**0.635**	**0.636**	**0.634**
BERT (Integrated Classes)	0.696	0.694	0.696	0.687

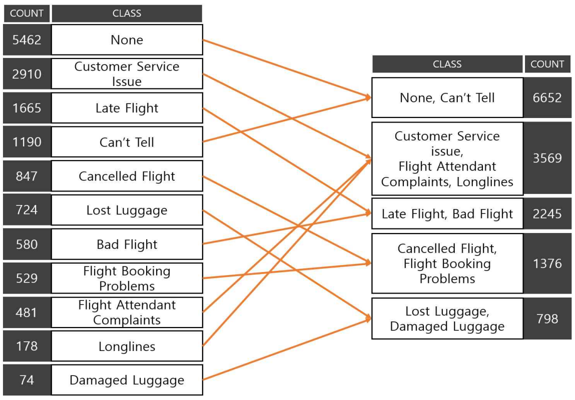


Fig. 2. Integrated Classes Relationships

$$Precision = \frac{TP}{TP+FP}, \quad (1)$$

$$Recall = \frac{TP}{TP+FN}, \quad (2)$$

$$F1\,score = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \quad (3)$$

The result of further learning of sentences made for negative reasons are marked as Additional Sentences. The BERT model, which has undergone further training on negative reason sentences, has shown the best results figures. However, it can be seen that the accuracy of the class with a small number of data is significantly lower. This is because the number of data in a particular class is small. To prove this, similar classes were integrated to make up for the number of scarce data. Fig. 2 visualized the integrated relationship of the class. The BERT (Integrated Classes) in Table 4 shows the complemented experimental results.

Ⅵ. Conclusion

In this work, negative reason has been extracted and classified from Tweet, unlike existing works that target sentimental classification. Various models have been applied, compared, and analyzed. Among them, after converting the target class into a sentence, the model that learned the added sentence showed the highest accuracy. This is a simple but powerful method in class classification similar to negative reason classification. It is expected that more accurate classification will be performed if the number of balanced data by class is obtained. Extracting such specific reasons for feedback can provide great value in many areas as well as airlines. Furthermore, these classification systems will provide the specific analysis results needed in areas such as public opinion analysis, disaster response, and will lead to faster response.

References

[1] A. Rane and A. Kumar, “Sentiment Classification System of Twitter Data for us Airline Service Analysis,” in 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), Vol. 1, pp. 769 - 773, IEEE, 2018.

[2] Q. Le and T. Mikolov, “Distributed Representations of Sen-tences and Documents,” in International conference on ma-chine learning, pp. 1188 - 1196, PMLR, 2014.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, “Bert: Pre-training of Deep Bidirectional Transformers for Language Under-standing,” arXiv preprint arXiv:1810.04805, 2018.

[4] A. Natekin and A. Knoll, “Gradient Boosting Machines, a Tu-torial,” Frontiers in neurorobotics, Vol. 7, pp. 21, 2013.

[5] D. Chen, J. Bolton, C. D. Manning, “A Thorough Examina-tion of the cnn/daily Mail Reading Comprehension Task,” arXiv preprint arXiv:1606.02858, 2016.

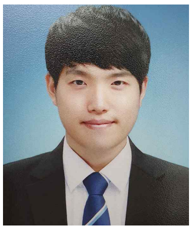
[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, “Attention is all you need,” arXiv preprint arXiv:1706.03762, 2017.

[7] Mikolov, T., Chen, K., Corrado, G., Dean, J., “Efficient Estimation of word Representations in Vector Space.” arXiv preprint arXiv:1301.3781, 2013.

[8] T. Chen and C. Guestrin, “Xgboost: A Scalable Tree Boosting System,” in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785 - 794, 2016.

[9] S. Hochreiter and J. Schmidhuber, “Long Short-term Memory,” Neural computation, Vol. 9, No. 8, pp. 1735 - 1780, 1997.

Jeong-Woo Guk (국 정 우)



2003 Computer Engineerning from Inha University (B.S.)
2009 Samsung SDS (Research & Development)
2019~Computer Science from Yonsei University (M.S)

Field of Interest: Machine Learning & Deep Learning & National Language Processing
Email: jw.kook@yonsei.ac.kr