Multi-label Text Classification of Economic Concepts from Economic News Articles using Natural Language Processing

Soojeong Kim
School of Electrical Engineering
Korea University
Seoul, South Korea
sookelly@korea.ac.kr

Minhyeok Lee
School of Electrical and Electronics Engineering
Chung-Ang University
Seoul, Korea
mlee@cau.ac.kr

Junhee Seok
School of Electrical Engineering
Korea University
Seoul, South Korea
jseok14@korea.ac.kr

***Abstract*—**Multi-label classification is rapidly developing as an important aspect of modern predictive modeling. In this paper, we propose a multi-label text classification approach in order to extract the labels of economic concepts from economic news articles. We demonstrate a multi-label sentence-level event classification with a multi-label classifier algorithm. The classifier uses BERT Model and classification based on the association between labels via a threshold. The experiment on real-world multi-label data with many labels demonstrates an appealing performance and efficiency of multi-label classification.

Keywords—Multi-label Classification, Natural Language Processing, Text Classification

I. INTRODUCTION

With a continuous increase in available data, there is a pressing need to organize it and modern classification problems often involve the prediction of multiple labels simultaneously associated with a single instance. Multi-label classification of textual data is an important problem. For instance, this can be employed to find the genres that a movie belongs to, based on the summary of its plot. [1] Multi-Label Classification is the supervised learning problem where an instance may be associated with multiple labels. This is an extension of single-label classification (i.e., multi-class, or binary) where each instance is only associated with a single class label. Unlike normal classification tasks where class labels are mutually exclusive, multi-label classification requires specialized machine learning algorithms that support predicting multiple mutually non-exclusive classes or labels.

In the economic domain, text classification tasks are highly popular for making available fundamental knowledge present in the economic text, such as business news. Categorizing economic concepts extracted from financial and business news can become a source of insights for enterprises. The domain of document classification is highly productive and general, and data are freely available, but economically focused resources are lacking. We use a multi-label classification method because each news item has one or more economic concepts. Multi-label classification of the textual data from the domain of economics, which is the focus of our paper, has not received much systematic attention so far. This work fills the shortcomings of economic and financial text mining applications.

II. PRELIMINARIES

A. Multi-label Classification

Multi-label classification problems are quite common in the real world. It is a challenging research problem that emerges in several modern applications such as music categorization [2, 3], bioinformatics such as protein function prediction [4], and semantic classification of images [5, 6]. Modern classification problems often involve the prediction of multiple labels simultaneously associated with a single instance, e.g., image tagging by predicting multiple objects in an image.

Multi-label classification can be performed in two different ways: problem transformation methods and algorithm adaptation methods. Problem transformation methods [7] transform the multi-label classification task into one or more single-label classification, regression, or label ranking tasks. Algorithm adaptation methods extend specific learning algorithms [8] in order to handle multi-label data directly.

B. Multi-label Text Classification

Traditional single-label classification is concerned with learning from a set of examples that are associated with a single label λ from a set of disjoint labels L, |L| > 1. If |L| = 2, then the learning task is called binary classification, while if |L| > 2, then it is called multi-class classification. In multi-label classification, the examples are associated with a set of labels $Y \subseteq L$. In the past, multi-label classification has mainly engaged the attention of researchers working on text categorization, as each member of a document collection usually belongs to more than one semantic category.

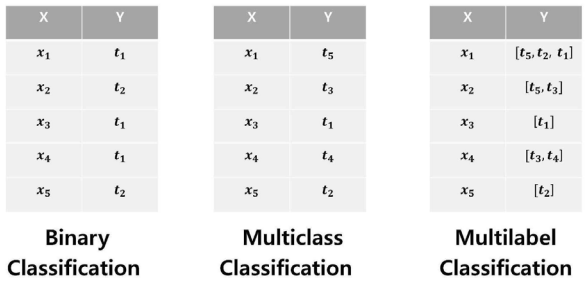

Fig. 1. Differences in Classification Tasks

One of the most used capabilities of supervised machine learning techniques is for classifying content, employed in many contexts like telling if a given restaurant review is positive or negative or inferring if there is a cat or a dog on an image. This task may be divided into three domains, binary classification, multiclass classification, and multi-label classification. Binary classification is used when there are only two distinct classes and the data to classify belongs exclusively to one of those classes, e.g., to classify if a text about a given target is positive or negative. Multi-class classification is used when there are three or more classes and

the data to classify belongs exclusively to one of those classes, e.g., to classify if an animal on an image is a dog, cat, or cow. Multi-label classification is used when there are two or more classes and the data to classify may belong to none of the classes or all of them at the same time, e.g., to classify which traffic signs are contained on an image.

III. PROPOSED METHOD

In this paper, we propose a multi-label sentence-level event classification algorithm. The main task is to predict labels based on economic concepts from financial and business news text presented in the actual dataset. In order to evaluate the proposed approach, we have divided the available data into training and testing samples for supervised learning. Each sample has a sentence and its true label vector. We use the pre-trained BERT model and fine-tune it for our classification task. We add an additional single dense layer on top of the pre-trained BERT model and the final hidden vector of the classification token [CLS] is fed into this dense layer. The special [CLS] token stands for 'classification' and will contain an embedding for the sentence-level representation of the sequence. We load the pre-trained model and then train the last layer for the classification task. The output from this task for each sentence is a meaningful list with classification scores for prediction labels.

For selecting strong prediction labels, we set a threshold value as the mean value of the classification score for the given sentence. It is easy to assume the classification threshold to be 0.5, however machine learning thresholds should be problem-specific. The normal default threshold value might not be the best way to understand the anticipated probability. The distribution of classes on economic concepts from news articles may be substantially skewed. For those classification problems that have a severe class imbalance, the default threshold can result in poor performance. As such, a simple and straightforward approach to improving the performance of a classifier that predicts probabilities on an imbalanced classification problem is to tune the threshold used to map probabilities to class labels. As we consider testing samples achieving a higher probability of predicted labels should be classified as positive, we need to roll it over to 1 or 0 depending on whether it is above or below the threshold value. The detailed procedure is shown in Fig.3 as pseudocode for the classifier algorithm. Fig.2 describes the architecture of the proposed multi-label sentence-level event classification model.

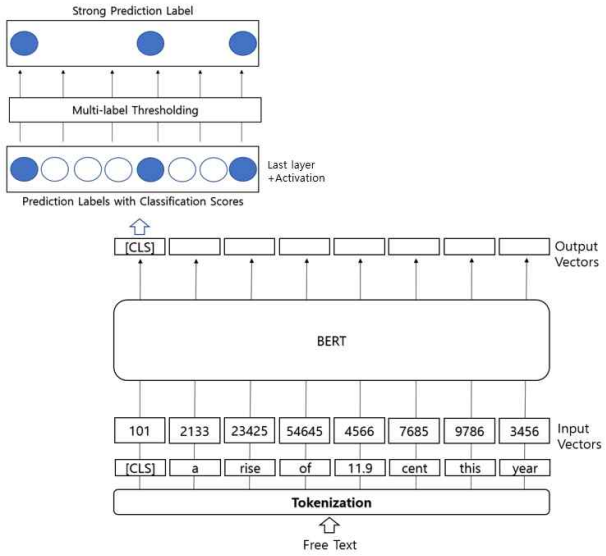

Fig. 2. The architecture of the proposed multi-label sentence-level event classification model

Algorithm 1: Multi-label Classifier Algorithm

Input: Training and testing samples of S, L(S: sequence of sentences, L: sequence of labels for each sentence)
Output: prediction label vector list γ
Load the pre-trained BERT model.
Tokenize and encode input S, L of training and testing samples
Train the pre-trained BERT model with encoded input vectors
Get classification score vector list by classifying S, L of testing samples with pre-trained BERT model
$\gamma = 0$
for each sequence of S in testing samples **do**:
  Fetch corresponding score vector σ *from classification score vector list*;
  Set threshold as mean of score values in σ;
  for each score value in score vector σ *with respect to true labels* **do** :
    if score value $\geq$ *threshold* :
      Set prediction label value as positive;
      Append prediction label value into prediction label vector;
    else:
      Set prediction label value as negative;
  End for
  $\gamma = \gamma$ *U prediction label vector*
End for
Return prediction label vector list γ

Fig. 3. Pseudocode for the Multi-label Classifier Algorithm

IV. EXPERIMENT RESULTS

We used the SENTiVENT event dataset [9] (3072 sentences, 18 labels), a corpus of English economic news articles on all companies in the S&P500 from various sources over a period of 14 months. This dataset focuses on annotating certain company-specific events in economic and financial news, each of which is related to extracting a certain economic concept. In this dataset, there are 18 unique attributes, which are 'CSR/Brand', 'Deal', 'Dividend', 'Employment', 'Expense', 'Facility', 'FinancialReport', ''Financing', 'Investment', ''Legal', 'Macroeconomics', 'Merger/Acquisition', ''Product/Service', 'Profit/Loss', 'Rating', 'Revenue', 'SalesVolume', and 'SecurityValue'. We split the core dataset into 2764 training sample and 308 testing sample.

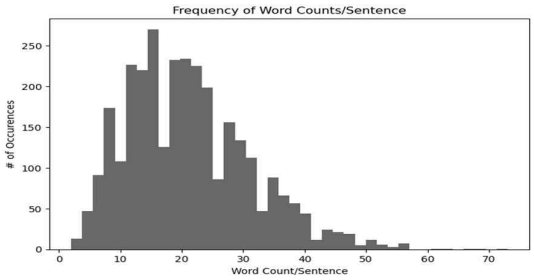

Fig. 4. Frequency of Word Counts per sentence

Transformer models such as BERT [10] cannot manage more than 512 words at a time. Its max length of tokens is limited to 512. A histogram plot shown in Fig. 4 reveals that most of the sentences have a word count under 50. Also, in general, that much length is reasonable for the model to develop sufficient context to be able to perform classification for a narrow problem. We will restrict ourselves to the first 50 words. A max word count of 50 seems reasonable since that should cover the text of most sentences.

We adjusted the epoch for the BERT experiment to confirm that the overfitting point is epoch 12. Fig. 5 describes the train and validation loss of the following procedure. The train and validation loss are continuously and rapidly reduced.

Fig. 5. Train and Validation Loss Graph

	Precision	Recall	F1-score
0	0.95	0.58	0.72
1	0.07	0.50	0.12
Accuracy			0.58
Macro avg	0.51	0.54	0.42
Weighted avg	0.90	0.58	0.69

Fig. 6. Classification Report of True and Predicted labels

In order to evaluate the performance of the multi-label classifier, a classification report based on a confusion matrix is used. The classification report of ground truth (correct) target values and estimated targets as returned by a classifier is shown in Fig.6. The classification report compares predictions we have made for the target variable with the real classes. The prediction shows 58%, 51%, and 54% each for accuracy, precision, and recall on average. The macro F1 score is 42% which calculates the F1 separated by class but not using weights for aggregation. The weighted F1 score is 69%, which calculates the F1 score for each class independently but when it adds them together it uses a weight that depends on the number of true labels of each class.

$$Exact\ Match\ Ratio, EMR = \frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i) \quad (1)$$

Another way to compute the accuracy is defined in (1). It is a trivial way to just ignore partially correct (consider them incorrect) and extend the accuracy considering used in single-label cases for multi-label prediction. This is called Exact Match Ratio, which is useful to measure completely correct prediction results. We used (1) to compare the accuracy of when using our fine-tuned threshold value and just a default one. The comparison of accuracy result for (1) when the threshold value is the mean value of the classification score for each given sentence and when the value is a default fixed value of 0.5 in the experiment is shown in TABLE I. It is the result showing the difference of accuracy of multi-label classification according to the threshold value. Fig.7 shows an example of true labels and prediction labels for a given sentence.

TABLE I. THE ACCURACY RESULT OF EXPERIMENT

	Experiment with 308 text sequences	
	Exact Match Ratio (threshold=mean value of each classification score)	***Exact Match Ratio (threshold=0.5)***
BERT	0.5	0.23

Sentence	However, the sell-off seems overdone as AMD's performance didn't deviate much from expectations.
True Label	'Financial Report', 'Security Value'
Prediction Label	'CSR/Brand', 'Deal', 'Expense', 'Financial Report', 'Legal', 'Macroeconomics', 'Merger/Acquisition', 'Product/Service', 'Security Value'

Fig. 7. Example of sentence, its true labels and prediction labels

V. CONCLUSIONS

The problem addressed in this paper is about extracting multi-labels of economic concepts from company-specific news articles. It makes use of document content and labels text to learn the label-specific document representation with the aid of a self-attention mechanism. We believe this experiment further enhances training models for advanced tasks like Relation Classification and NamedEntityRecognition (NER). Recently, there's been a surge in the popularity of various NLP models for classification tasks such as XLNet [11], ERNIE [12], Text-to-Text Transfer Transformer (T5) [13]. We are certain that such studies can incorporate into further multi-label classification research.

ACKNOWLEDGEMENT

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1F1A1050977) and (NRF-2022R1A2C2004003) Correspondence should be addressed to jseok14@korea.ac.kr.

REFERENCES

[1] Rafael B. Mangolin, Rodolfo M. Pereira, Alceu S. Britto Jr., Carlos N. Silla Jr., Valéria D. Feltrim, Diego Bertolini, Yandre M. G. Costa, "A multimodal approach for multi-label movie genre classification", 2020.

[2] Sergio Oramas, Oriol Nieto, Francesco Barbieri, Xavier Serra, "Multi-label Music Genre Classification from Audio, Text, and Images Using Deep Features" In Proceedings of the 18th International Society of Music Information Retrieval Conference(ISMIR), 2017.

[3] Michael A Casey, Remco Veltkamp, Masataka Goto, Marc Leman, Christophe Rhodes, and Malcolm Slaney, "Content-based music information retrieval: Current directions and future challenges", Proceedings of the IEEE, 2008.

[4] Dan Ofer, Nadav Brandes, Michal Linial, "The language of proteins: NLP, machine learning & protein sequences", Computational and Structural Biotechnology Journal, 2021.

[5] Fengtao Zhou, Sheng Huang, Yun Xing, "Deep Semantic Dictionary Learning for Multi-label Image Classification", AAAI, 2021.

[6] Fengtao Zhou, Sheng Huang, Yun Xing, "Deep Semantic Dictionary Learning for Multi-label Image Classification",

[7] Zhang, J.; Wu, Q.; Shen, C.; Zhang, J.; and Lu, J. 2018a. Multilabel image classification with regional latent semantic dependencies. IEEE Transactions on Multimedia, 2010.

[8] Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; and Xu, W, "CNN-RNN: A unified framework for multi-label image classification", In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[9] Gilles Jacobs, SENTiVENT Event Annotation Guidelines v1.1, Language and Translation Technology Team, Ghent University, https://osf.io/xqzw5, 2020.1.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" arXiv preprint arXiv:1810.04805, 2018.

[11] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding", 33rd Conference on Neural Information Processing Systems (NeurIPS), 2019.

[12] Zhang, Zhengyan and Han, Xu and Liu, Zhiyuan and Jiang, Xin and Sun, Maosong and Liu, Qun, "ERNIE: Enhanced Language Representation with Informative Entities", Proceedings of ACL, 2019.

[13] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", Journal of Machine Learning Research, 2020.