MULTI-LABEL INDEXING TECHNOLOGY FOR NEWS WITH AI-BASED TEXT PROCESSING

Y. Yasuda[1], S. C. Clippingdale[2], T. Miyazaki[1], J. Goto[1]

[1]NHK (Japan Broadcasting Corporation), Japan and [2]NHK Foundation, Japan

ABSTRACT

Broadcast media organisations produce many news scripts every day for dissemination as content. Such text data is often reused in the process of producing TV programmes and web news. To efficiently utilise this much data, it is necessary to accurately attach metadata such as labels that indicate the content of the text. However, manually assigning labels takes an enormous amount of time and effort. With the aim of reducing costs, we have developed a system that automatically labels news articles. A major challenge in the multi-label text classification task in the news domain is known as 'imbalanced learning.' We proposed a novel loss function that utilises some weights and a label-smoothing technique to suppress label imbalance. Experimental results show that our method outperforms baselines. We introduce a prototype system based on our method as a test bed for content creation and discuss some of the results that it achieves.

INTRODUCTION

Much text data is utilised in the process of producing TV programmes and web news. To create media content efficiently, it is necessary to attach accurate metadata, such as labels that indicate the content, to large amounts of text. Metadata attached to text can enable producers to efficiently retrieve and use past material in the creation of new content and enable viewers to easily access articles that they want.

Conventionally, metadata indicating content has been added to text data manually. In recent years, the amount of text data that needs to be handled has grown very large, driving a need to develop technology that supports the task of metadata generation to reduce costs, particularly for producers and broadcast stations with limited numbers of staff.

We have developed a system that uses AI technology to automatically assign multiple labels representing genre and content to news articles. This system performs multi-label text classification based on neural networks and makes it possible to add accurate metadata to large amounts of text in less time than is required for conventional manual metadata addition. The system was introduced and tested at two local broadcast stations. In one instance it was used to analyse the topics of news articles over a one-year period, and in the other instance, to add labels to news articles published online.

News articles cover a wide variety of topics in general, and the system must assign multiple labels to them. Since the labels vary from major genre areas down to quite detailed minor topics, their frequencies of appearance vary widely, and the labels for minor

topics in the training data can appear extremely infrequently. It is known that when AI models are trained with datasets including labels that occur infrequently, the classification accuracy deteriorates. We therefore developed a learning algorithm that suppresses the influence of wide imbalances in label appearance frequencies and aimed to improve its performance.

MULTI-LABEL TEXT CLASSIFICATION

Multi-label text classification is a key task in natural language processing that is applied to various situations in the real world [1], [2], [3]. Examples of real-world usage include classification of legal documents [4] and automatic diagnosis through medical records [5]. In multi-label text classification, the task is to assign an appropriate label subset $l^{(i)}$ ($l^{(i)} \in L$) to document $x^{(i)}$ ($x^{(i)} \in X$), where L is the set of predefined labels, X is the set of documents and i is the index of input samples. Fig. 1 shows the conceptual image of multi-label text classification.

Following rapid advances in machine learning technology in recent years, multi-label text classification is often tackled with neural networks. We also approach this task using a neural network paradigm.

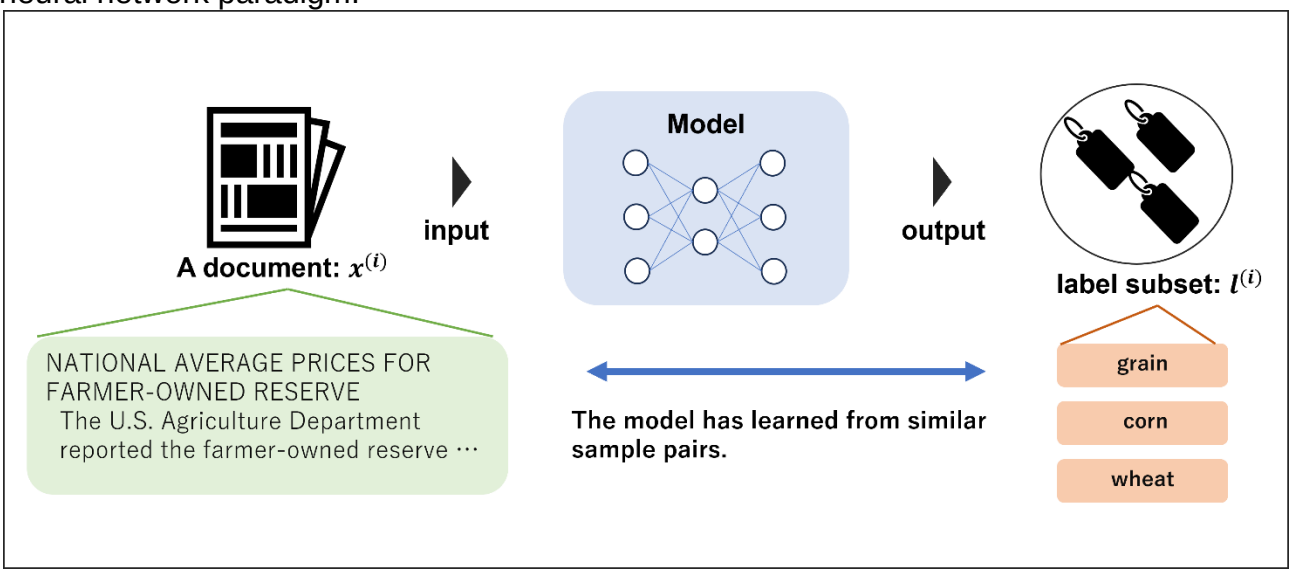

Figure 1 – Conceptual image of the multi-label text classification task

Learning of A Model in Multi-label Text Classification

The task of multi-label learning is to learn from training data a function $f: \chi^{(i)} \rightarrow y^{(i)}$, where $\chi^{(i)}$ ($\chi^{(i)} \subset \mathbb{R}^d$) denotes a d-dimensional feature vector representing a document and $y_{\square}^{(i)} = \left\{y_1^{(i)}, y_2^{(i)}, \ldots, y_N^{(i)}\right\} \in \{0,1\}^N$ denotes ground-truth values as a multi-hot vector of labels. A model outputs the estimated probability p_n of the n-th label, and the probabilities are fed into a loss function, which measures the loss as the difference between the model's output p_n and the ground-truth value $y_n^{(i)}$. For the sake of simplicity, we omit (i), the index of a sample, in the rest of this paper.

Binary cross entropy (BCE) is a widely used loss function in the field of multi-label text classification. BCE is defined for one sample as

$$\mathrm{BCE} = -\sum_{\mathrm{n}=1}^{\mathrm{N}} [y_n \log p_n + (1 - y_n) \log(1 - p_n)] \quad Eq.(1)$$

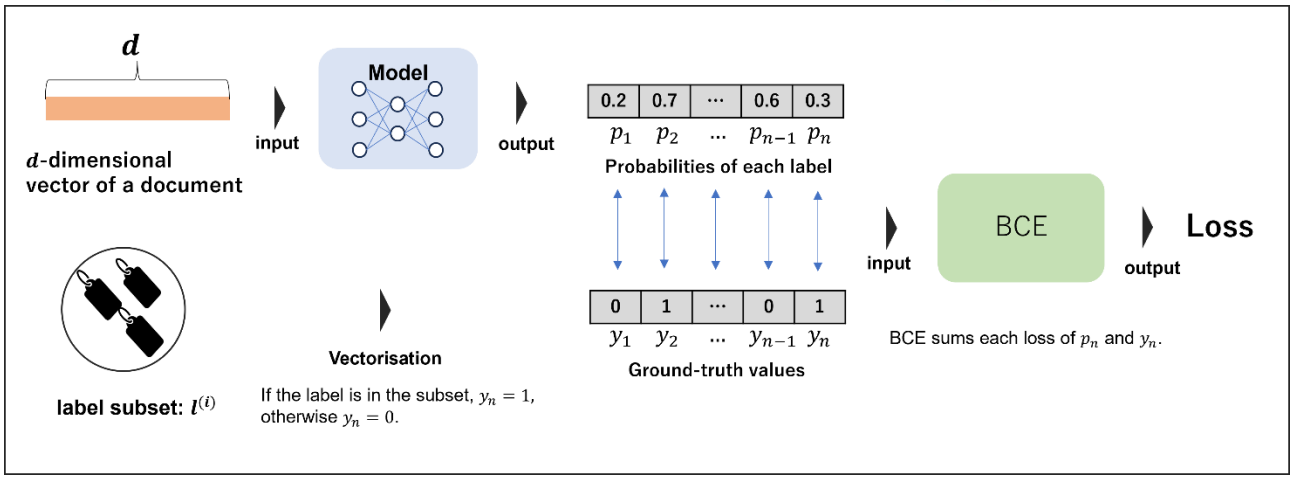

Figure 2 – Model learning with BCE loss in the multi-label text classification task.

Fig. 2 shows a conceptual image of the model learning with BCE loss for one sample. The probabilities p_n as outputs of the model and the ground-truth values y_n are input to the BCE computation. The ground-truth values y_n comprise a multi-hot vector defined on the label subset. If the label corresponding to the n-th position is assigned to the input document, y_n becomes 1; otherwise it is 0. Therefore, if the label is positive for the sample, only the term $\log p_n$ is summed into the final loss value BCE in Eq. 1. On the other hand, if $y_n = 0$, the term $\log(1 - p_n)$ is summed into the final loss value. The terms $\log p_n$ and $\log(1 - p_n)$ represent the loss associated with differences between the probabilities produced by the model and the ground-truth values. BCE calculates the loss for each label and then calculates the sum for one sample. The model learns to reduce the loss by making the output probabilities $\{p_n\}$ closer to the ground truth values $\{y_n\}$.

The Challenge of the News-Domain Text Classification Task

In real-world applications, multi-label text classification is a very important technology for automatic metadata assignment to content. Generally, a document in the real world contains multiple concepts, so single-label text classification tasks are insufficient for relating documents using metadata. For example, simply labelling an article about a tennis match as 'sports' would not be a very useful metadata addition. Instead, we can organise the data more usefully by adding more detailed labels such as 'tennis' or 'Wimbledon' as metadata.

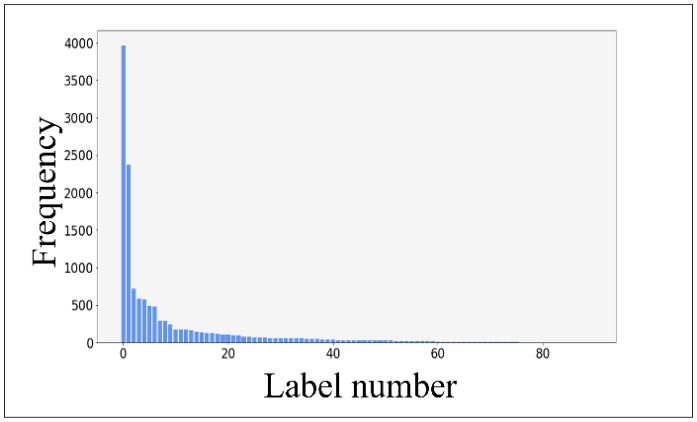

Figure 3 – Label distribution in Reuters-21578. The vertical axis represents the frequency of labels on the dataset, and the horizontal axis represents the index n assigned to the labels in frequency order.

In multi-label text classification, many predefined labels suitable for the domain could be constructed. Especially in the news domain, a very wide range of topics must be covered. However, it is difficult to train neural network models to output accurate labels on training data with a wide variety of detailed topics. This is because there are significant differences in the frequencies of occurrence of labels in the training data [6], [7]. As an example, Fig. 3 shows label frequencies for the news-based benchmark dataset Reuters-21578 for multi-label text classification. Reuters-

21578 consists of 21,578 documents taken from the original Reuters-22173 corpus after the removal of 595 duplicate documents by Lynch and Lewis in 1996 [8], [9], [10]. As the figure shows, the label frequencies follow a long-tailed distribution.

Such a long-tailed label distribution leads to a decrease in the accuracy of neural networks. Major (high-frequency) labels that appear frequently in the training data have many documents to train on, while minor (low-frequency) labels that appear rarely have a far smaller number of document types to train on. As a result, the neural network may become overtrained to specific document features and minor labels, and will often not output minor labels on documents that do not contain an exact match with the learned sentences or words. This is a major problem in multi-label text classification in general, but is particularly prevalent in the news domain due to the wide range of topics covered.

AUTOMATIC NEWS LABELLING SYSTEM

In this section, we describe a prototype Automatic News Labelling System developed using multi-label text classification. The system automatically assigns labels indicating the genre and content of news articles entered through a web-based interface. Fig. 4 shows the process flow of the assignment of labels to a news article.

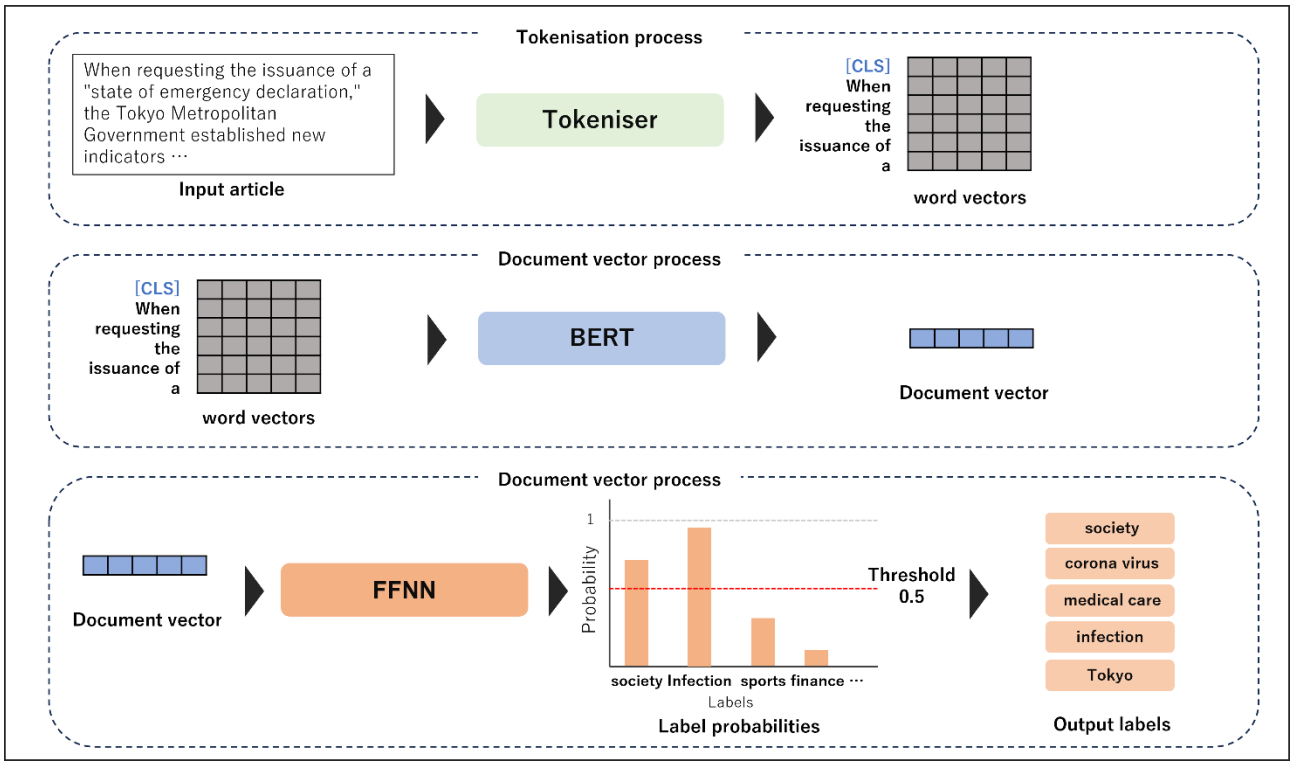

Figure 4 – Process flow in labelling a news article in our system

First, the article provided by the user is tokenised by the tokeniser. To process natural language using a neural network, it is necessary to segment the language into 'tokens.' This is because neural network models cannot process language *per se*, but use a dictionary to map short input word sequences (tokens) into corresponding vectors (i.e. a numerical representation) for processing. However, unlike English, where tokens may correspond to individual words, Japanese words are not separated by spaces, so we split sentences into tokens using other sentence parsing techniques [11].

The split tokens are then fed into a BERT ('Bidirectional Encoder Representations from Transformers') network [12]. BERT is a type of Transformer [13] neural network that uses attention techniques and is known to deliver stable performance in a variety of natural

language processing tasks. The BERT network in our system was pre-trained using NHK news articles and Twitter data. The first token output as a feature of BERT is determined to be a special token used for classification tasks. The system treats the output vector of this special token as a vector representing the meaning of the entire document and extracts it.

Finally, the system feeds the extracted vector of special tokens into a feed-forward neural network (FFNN). The vector dimensionality is fixed at 768 for BERT. However, the number of labels we want to classify is not necessarily 768. A FFNN is responsible for converting vectors that represent the meaning of a sentence into score vectors whose dimension is the number of labels we want to classify. The system calculates a score (probability) for each label and outputs those labels whose score exceeds a threshold (0.5 in this case).

Fine-tuning of the Model Utilised in the System

In recent years, most language models such as BERT used in various applications have been pre-trained to acquire general linguistic information using a large amount of text. However, for downstream tasks such as multi-label text classification, the model must be trained with more task-specific data ('fine-tuning') to adapt the model better to the task at hand.

In order to fine-tune the language model used in this system, we constructed a multi-label text classification dataset based on NHK news articles as the domain. Based on past articles published on NHK's web news site, NHK NEWS WEB, we assigned labels indicating the content to about 50,000 articles in the data set. The total number of labels in the dataset was 1,015, and the labels were defined as a set of nouns that can generically represent the content of the articles. All labels were determined by consensus between annotators and supervisors by reading published articles. Subsets of labels were assigned to articles according to consensus between two annotators.

Since we did not intentionally eliminate label imbalance in the dataset we constructed, the model in this system is also expected to have low accuracy for low-frequency labels. We therefore constructed a new loss function to reduce the effect of label imbalance.

Suppression of Effect of Label Imbalance

BCE is often used in multi-label text classification, but it is vulnerable to label imbalance effects. The ultimate loss value for one sample is the sum of the loss values from each label. But across the whole dataset, most of the loss from low-frequency labels is dominated by the terms $\log(1 - p_n)$ in Eq. (1) contributed by the many negative samples (the many articles n for which the label is absent).

In addition, an imbalance between positive and negative labels in a single input sample may have a negative impact on model training. In news article classification tasks, a large number of labels are predefined. For example, in our prototype system, we predefined 1,015 labels $(N = 1015)$. On the other hand, a single news article tends to be assigned no more than 10 labels. The ultimate loss is constructed by the summation of the loss values from all (1015) labels, so a large proportion of the loss for a single sample is dominated by losses from the negative labels. The imbalance disturbs meaningful learning of the positive labels for which the model should learn to output higher probabilities.

In this paper, we propose a loss function that combines a weight based on label frequencies, a weight that reduces the influence of negative samples, and a weight based on label co-occurrence information. This loss function aims to suppress the impact of imbalanced datasets including many low-frequency labels on the training of a model.

We propose a method called weighted asymmetric loss (WASL), inspired by the asymmetric loss (ASL) [14] that has been proposed in the image classification domain. We aim to appropriately select loss values from negative labels to suppress, and we adjust the suppression in accordance with the label frequencies. Furthermore, by performing label smoothing (LS) [15] based on co-occurrence of (correlations between) the labels, we compensate for the small number of samples of low-frequency labels and suppress the overfitting of the model. We define WASL as

$$\mathrm{WASL} = -\sum_{\mathrm{n=1}}^{\mathrm{N}} w_n [y'_n \log L_+ + (1 - y'_n) \log L_-] \quad Eq.(2)$$

$$L_+ = (1 - p_n)^{\gamma_+} \log p_n \quad Eq.(3)$$

$$L_- = (p'_n)^{\gamma_-'^{(n)}} \log(1 - p'_n) \quad Eq.(4)$$

$$p'_n = \max(p_n - \mathrm{m}, 0) \quad Eq.(5)$$

$$\gamma_-'^{(n)} = w_n \gamma_-, \qquad Eq.(6)$$

$$w_n = \frac{\frac{(1-\beta)}{1-\beta^{c_n}} N}{\sum_n^N \frac{(1-\beta)}{1-\beta^{c_n}}} \quad Eq.(7)$$

where β ($\beta \in [0,1]$) is a hyperparameter that determines the strength of the class-balanced weights and m is a hyperparameter that is sufficiently small to discard the loss values from negative samples for which learning has sufficiently progressed. γ_+ and γ_- $(0 \leq \gamma_+ < \gamma_-)$ are hyperparameters that adjust the balance between loss values from positive or negative labels. c_n represents the frequency of the n-th label in the dataset.

The three principal concepts underlying our proposed method are as follows: suppression of loss values from negative labels by $\gamma_-'^{(n)}$; re-sampling of label frequencies by w_n; and the use of the smoothed ground-truth values y'_n in Eq. (2) in place of the original y_n (see below).

In multi-label text classification, for any given sample (article), most labels are negative, meaning that most of the y_n are 0. As a result, the total loss value is dominated by $(1 - y'_n) \log L_-$ in Eq. (2), representing the loss derived from negative labels. During the training process, the model can reduce the loss value by making the output probability of the corresponding label closer to the ground-truth value. This dominance of negative labels prevents the model from learning meaningfully. Thus, we introduce in our method the weight $\gamma_-'^{(n)}$ in Eq. (4). $\gamma_-'^{(n)}$ is determined by the hyperparameter γ_- and the weight w_n to adjust for the effect of label frequency. It can selectively reduce the loss values derived from already well-learned negative labels.

We introduce the class-balanced weight utilised in class-balanced loss (CBL) w_n in Eq. (7) so that the model can more appropriately consider the differences in label frequencies [16]. As a label y_n appears more frequently in a dataset, w_n becomes smaller. This reduces loss values associated with high-frequency labels and increases those associated with low-frequency labels. CBL was proposed by Cui et al. [16] for correcting label imbalance in a

single-label classification task but has also proven successful in text classification (17) so in this work, we use this extension of CBL for multi-label text classification.

Label smoothing (LS) techniques proposed by Szegedy *et al* [15] can suppress overfitting of a model and calibrate the model. We perform LS with the label co-occurrence information set as prior probabilities to suppress overfitting of the model. LS here is defined as

$$y'_n = (1-\alpha)y_n + \alpha\ \mathrm{norm}(o_n)\quad Eq.(8)$$

$$\boldsymbol{o} = \boldsymbol{y}\cdot\boldsymbol{P}\quad Eq.(9)$$

where α ($\alpha \in [0,1]$) is a hyperparameter that determines the degree of smoothing and $\mathrm{norm}(\cdot)$ is the min-max normalisation function. Also, $\boldsymbol{y}$ ($y_n \in \{0,1\}^N$) is a multi-hot vector for the input sample and $\boldsymbol{P}$ indicates a square matrix of positive pointwise mutual information (PPMI) calculated from the number of co-occurrences of each label [18], [19]. Fig. 5 shows an example smoothed target vector produced by our method, where the labels 'wheat', 'flour', and 'grain' tend to co-occur with each other in the training data. If 'wheat' and 'grain' are positive labels for some input sample, that sample is likely to also be associated with the label 'flour.' Some value is re-distributed to related labels for some sample (from 'wheat' and 'grain' to 'flour' in this case) according to the strength of the co-occurrence (PPMI score) across the whole dataset. On the other hand, no value is distributed to the label 'gold' because it does not co-occur with 'wheat' and 'grain' in the training data. The model can learn from input sample - relevant label pairs, even if the labels are not truly positive labels.

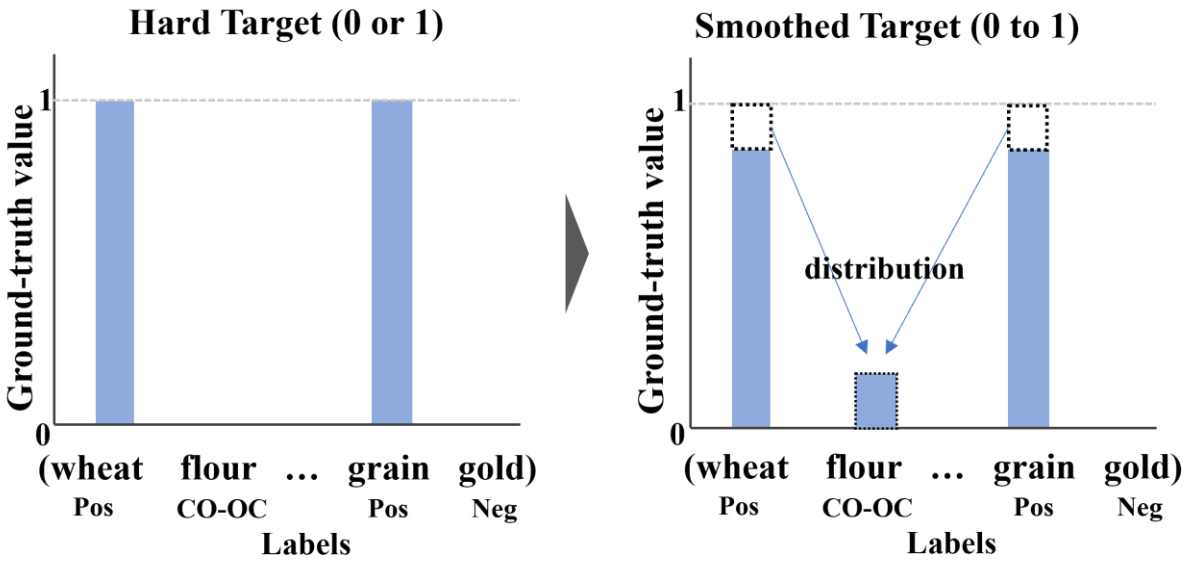

Figure 5. – Hard target $\boldsymbol{y}$ (multi-hot vector) and smoothed target $\boldsymbol{y}'$ produced by label smoothing in our method.

EXPERIMENT

We conducted a comparative experiment with baseline methods to investigate the effectiveness of the proposed method. We used macro-f1 and micro-f1 as evaluation metrics. Reuters-21578 and NHK NEWS WEB were utilised in this experiment as benchmark datasets. The model was implemented using PyTorch [20] and Transformers [21], and the optimisation method was AdamW [22]. The common hyperparameters in both datasets and all methods were as follows. The number of learning epochs was set to 50, dropout rate was 0.1, and output threshold was unified to 0.5. All hyperparameters (except

the common ones) were determined by using grid search according to micro-f1 on validation data. Table 1 lists the results of this experiment. As we can see in Table 1, the proposed method significantly outperformed the baselines on both datasets. These results suggest that the proposed method is effective for training a model on an imbalanced dataset.

Dataset Methods / Metrics	Reuters-21578 Macro-f1 / Micro-f1	NHK NEWS WEB Macro-f1 / Micro-f1
BCE	0.586 / 0.903	0.488 / 0.732
ASL	0.644 / 0.904	0.503 / 0.745
WASL (proposed)	**0.669** / **0.911**	**0.506** / **0.754**

Table 1 - Experimental results. Means of five trials are shown for the proposed method and baselines

AUXILIARY MODULES OTHER THAN AUTOMATIC LABELLING

In addition to the automatic labelling function, the system also contains modules for extracting keywords and listing related articles. These modules are implemented with the aim of facilitating content creation assistance based on automatic labelling.

The keyword extraction function highlights place names and proper nouns in a news article provided by the user. This can assist the user in creating and applying new fine-grained labels. In multi-label text classification using neural networks, all the labels to be assigned must be defined in advance, and it can be difficult to add topical news article labels corresponding to the latest popular words or place names. However, there is a considerable need among content producers for such functions that can add metadata such as labels for proper nouns and place names. By utilising named entity extraction technology, likely proper nouns and place names in the text can be extracted, listed, and highlighted to assist producers in considering the creation of new labels.

The system can also list published articles in a news database that are related to an article input or specified by the user. Some news topics have a high degree of continuity from previous articles, and when publishing new articles, including links to related prior articles increases reader satisfaction and engagement.

This module is implemented by comparing vectorised news articles using cosine similarity scores. By using cosine similarity, we eliminate the effect of the indeterminate number of labels assigned to a news article, making it possible to measure the relevance of an article according to its content.

Articles in the database have labels like those assigned to the input article by the system. The system can automatically label large amounts of old article data that does not yet have metadata, making it easier to reuse large amounts of old articles that have not thus far been organised for content creation.

USE CASES

NHK has local broadcasting stations in each Japanese prefecture providing content that is specialised for each region. Such stations have fewer staff and need more efficient content production, so we introduced our prototype Automatic News Labelling System in related but different use cases at two local stations.

Programme Creation Reusing News Articles

Here we explain an example where the system was used for content reuse. Broadcasters often need to reuse published content to help create new content. However, it may be the case that the content they have created does not have the metadata necessary for reuse.

At the end of 2023, producers at a local broadcast station wished to create a segment in a news programme about ranking news articles published on the local news website over the course of a year by topic. Unfortunately, the news articles published over the one-year period were not annotated with topic metadata. We used our system to automatically assign labels and re-aggregate the published article data by topic. Specifically, we reaggregated data from 121 articles, and the labels automatically assigned by the system were manually reviewed and confirmed. The rankings were then shown on news programmes to help viewers visually grasp local events from the past year.

Automatic Labelling of News Articles

In NHK, each regional station publishes local news on its website for the prefecture it serves. Unlike the national news site, the local news sites did not label news articles due to limited staffing. It is hoped that the introduction of our system will allow efficient metadata addition even with a small number of staff, enabling viewers also to access such information associated with local services.

In the article publishing workflow, producers input completed pre-publication news articles into the system. They then check and correct the output labels, and manually enter new labels if necessary. Labels constructed in this way are then published on the website alongside the news articles and displayed as hyperlinks to related articles.

Operational Issues

We found that using this system can help broadcasters with limited staff to produce rich content. However, some practical issues were also identified.

News articles on local news websites tend to be focused on unique traits of the region, and producers in local regions want more specific labels expressing these unique regional features. The labels defined in this system, however, are intentionally constructed to be common words. To attract the interest of local viewers, more specific proper nouns should be assigned to news articles instead of generic words. It is difficult to collect a large number of labels that represent the characteristics of the region. Even if we collect a few samples, such labels are trained even less frequently than other minor labels, making it difficult to solve this problem using only the method in this paper. Therefore, it will be necessary to prototype new algorithms, such as generative methods.

The user interface of the system was also identified as a major practical issue. Our system was initially conceived and intended to be usable nationally by NHK, and we built the interactive interface accordingly. However, in practice, producers needed to deal with many articles at once, for example for aggregation work. When using the system for content reuse, it was reported that the system's interface was not a good fit with the production workflow, resulting in inefficient use of time. We thus concluded that we should implement an interface more suitable for individual workflows, including for example a batch processing interface.

RELATED WORK

Many studies have proposed model architectures to improve the quality of multi-label classification. For example, Chen *et al* [23] developed a basic encoder-decoder model using a convolutional neural network and recurrent neural network for multi-label classification. Adhikari *et al* [24] reported that a well-tuned simple bidirectional-LSTM model can outperform some complex models. Models considering the relationship between labels have also been proposed. Yang *et al* [25] proposed a sequence generation model that treats the multi-label text classification task as a sequence generation problem with the aim of considering label relations. Also, Xiao *et al* [26] achieved a state-of-the-art performance on the AAPD dataset of Yang *et al* [25]. While many methods focusing on architecture have been proposed, such complex approaches tend to require extensive computing resources. Furthermore, even simple pre-trained models like BERT [12] still maintain a state-of-the-art performance on some datasets. Most existing models for multi-label text classification are trained with BCE, which makes them susceptible to label distribution, especially for imbalanced learning. Therefore, an approach that focuses on the loss function is required.

From a different perspective, methods that utilise dependencies between labels can also help improve the accuracy on multi-label text classification tasks. Dembczyński *et al* [27] pointed out the importance of considering label correlations in the multi-label text classification domain. Pal *et al* [28] proposed a model that treats the relationships between labels by means of graph attention networks [29]. Zhao *et al* [30] developed a model that creates clusters of labels and extracts the correlations between clusters. Song *et al* [31] achieved state-of-the-art performance on some datasets with a method combining a cloze task and multi-label text classification. On the other hand, many methods that take into account the relationship between labels have been proposed in the image classification domain [32], [33], [34]. Many methods based on label correlation utilise label embedding techniques [35] and graph convolutional networks [36]. Methods based on adjusting ground-truth values directly, for example LS, are not discussed so much. LS can suppress overfitting of a model and calibrate the model [37]. We believe that LS can be helpful for improving the accuracy of models for multi-label text classification.

CONCLUSIONS

Broadcasting stations produce many news scripts every day for TV programmes and web content. In order to efficiently use and re-use this large amount of text data, metadata labels that indicate the contents of the manuscripts must be accurately assigned. However, manually labelling takes a huge amount of time and effort, and in order to improve efficiency and significantly reduce the burden and cost of news production, we developed a prototype system that uses AI to automatically label news manuscripts.

To realise this system, we utilise multi-label classification, a task in which a computer outputs the applicable subset of labels for a given input manuscript. The model learns from training pairs consisting of an article and its associated label subset.

Since news articles cover a vast range of topics, label classification requires an enormous number of words to be prepared as label candidates, but in most cases only a few appropriate labels are assigned to any single article. As a result, previous research has shown that the adjustment of probabilities during learning is biased towards the process of reducing the probability of labels that do not match the content of the article being wrongly output (rather than increasing the probability of those labels that do appear being correctly output). This has a negative impact on learning and leads to a decline in classification performance.

To tackle this issue, we developed a novel loss function to suppress the effects of this 'label imbalance'. The loss function we developed utilises some weights that reduce the loss values from negative labels and high-frequency labels. Additionally, label smoothing based on label co-occurrences is introduced to suppress overfitting of the model to low-frequency labels. Evaluation experiments confirmed that our method improved classification performance on imbalanced data.

The prototype Automatic News Labelling System we developed has been introduced for testing at two local broadcasting stations and has achieved a reduction in the effort required for content production, while identifying some workflow issues that remain to be tackled. We expect subsequent versions of this and other similar systems to allow producers to create richer content in the future with less cost and effort.

REFERENCES

1. Zhang, M.-L. and Zhou, Z.-H., 2014. A Review on Multi-Label Learning Algorithms. IEEE Transactions on Knowledge and Data Engineering. Vol.26, No.8, pp. 1819 to 1837.
2. Tsoumakas, G. and Katakis, I., 2007. Multi-Label Classification: An Overview. International Journal of Data Warehousing and Mining, Vol.3, No.3, pp. 1 to 13.
3. Ueda, N. and Saito, K., 2002. Parametric Mixture Models for Multi-Mabeled Text. Advances in Neural Information Processing Systems. Vol.15.
4. Chalkidis, I., Fergadiotis, E., Malakasiotis, P., Aletras, N., and Androutsopoulos, I., 2019. I. Extreme Multi-Label Legal Text Classification: A case study in EU Legislation. In Proceedings of the Natural Legal Language Processing Workshop 2019. pp. 78 to 87.
5. Yao, L., Mao, C., and Luo, Y., 2019. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC Medical Informatics and Decision Making, Vol.19, No.3, pp. 31 to 39.
6. He, H. and Garcia, E. A., 2009. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, Vol.21, No.9, pp. 1263 to 1284.
7. Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M., 2020. Generalizing from a Few Examples: A Survey on Few-shot Learning. ACM Computing Surveys (csur), Vol.53, No.3, pp. 1 to 34.
8. Joachims, T., 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In Proceedings of the 10th European Conference on Machine Learning, pp. 137 to 142.
9. Lewis, D. D., 1992. An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 37 to 50.
10. Yang, Y. and Liu, X., 1999. A re-examination of text categorization methods. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 42 to 49.
11. Kudo, T. and Richardson, J., 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 66 to 71.

12. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol.1, pp. 4171 to 4186.

13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I., 2017. Attention is All you Need. Advances in Neural Information Processing Systems. Vol.30.

14. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., and Zelnik-Manor, L., 2021. Asymmetric Loss for Multi-Label Classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision. pp.82 to 91.

15. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z., 2016. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818 to 2826.

16. Cui, Y., Jia, M., Lin, T.-Y., Song, Y. and Belongie, S., 2019. Class-Balanced Loss Based on Effective Number of Samples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9268 to 9277.

17. Huang, Y., Giledereli, B., K¨oksal, A., ¨Ozg¨ur, A., and Ozkirimli, E., 2021. Balancing Methods for Multi-label Text Classification with Long-Tailed Class Distribution. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 8153 to 8161.

18. Church, K. W. and Hanks, P., 1990. Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics, Vol.16, No.1, pp. 22 to 29.

19. Niwa, Y. and Nitta, Y., 1994. Co-Occurrence Vectors From Corpora vs. Distance Vectors From Dictionaries. In COLING The 15th International Conference on Computational Linguistics. Vol.1.

20. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., and Antiga, L. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems, Vol. 32.

21. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38 to 45.

22. Loshchilov, I. and Hutter, F., 2019. Decoupled Weight Decay Regularization. In International Conference on Learning Representations.

23. Chen, G., Ye, D., Xing, Z., Chen, J., and Cambria, E., 2017. Ensemble Application of Convolutional and Recurrent Neural Networks for Multi-label Text Categorization. In 2017 International Joint Conference on Neural Networks, pp. 2377 to 2383.

24. Adhikari, A., Ram, A., Tang, R., and Lin, J., 2019. Rethinking Complex Neural Network Architectures for Document Classification. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol.1, pp. 4046 to 4051.

25.Yang, P., Sun, X., Li, W., Ma, S., Wu, W., and Wang, H., 2018. SGM: Sequence Generation Model for Multi-label Classification. In Proceedings of the 27th International Conference on Computational Linguistics, pp. 3915 to 3926.

26.Xiao, L., Huang, X., Chen, B., and Jing, L., 2019. Label-Specific Document Representation for Multi-Label Text Classification. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 466 to 475.

27.Dembczyński, K., Waegeman, W., Cheng, W., and Hüllermeier, E., 2012. On label dependence and loss minimization in multi-label classification. Machine Learning, Vol.88, pp. 5 to 45.

28.Pal, A., Selvakumar, M., and Sankarasubbu, M., 2020. Multi-Label Text Classification using Attention-based Graph Neural Network. arXiv preprint ArXiv:2003.11644.

29.Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. 2018. Graph Attention Networks. arXiv preprint arXiv:1710.10903.

30.Zhao, F., Ai, Q., Li, X., Wang, W., Gao, Q., and Liu, Y. 2024. TLC-XML: Transformer with Label Correlation for Extreme Multi-label Text Classification. Neural Processing Letters, Vol.56, No.1, pp. 1 to 25.

31.Song, R., Liu, Z., Chen, X., An, H., Zhang, Z., Wang, X., and Xu, H., 2023. Label prompt for multi-label text classification. Applied Intelligence, Vol.53, No.8, pp. 8761–8775.

32.Chen, Z.-M., Wei, X.-S., Wang, P., and Guo, Y., 2021. Learning Graph Convolutional Networks for Multi-Label Recognition and Applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.45, No.6, pp. 6969 to 6983.

33.Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., and Xu, W., 2016. CNN-RNN: A Unified Framework for Multi-Label Image Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2285 to 2294.

34.Ye, J., He, J., Peng, X., Wu, W., and Qiao, Y., 2020. Attention-Driven Dynamic Graph Convolutional Network for Multi-Label Image Recognition. In Proceedings of Computer Vision–ECCV 2020: 16th European Conference, pp. 649 to 665.

35.Zhang, H., Xiao, L., Chen, W., Wang, Y., and Jin, Y., 2018. Multi-Task Label Embedding for Text Classification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4545 to 4553.

36.Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I., and Welling, M., 2018. Modelling Relational Data with Graph Convolutional Networks. In Proceedings of the Semantic Web: 15th International Conference, pp. 593 to 607.

37.Müller, R., Kornblith, S., and Hinton, G. E., 2019. When Does Label Smoothing Help? Advances in Neural Information Processing Systems, Vol.32.