2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

An Empirical Study of Automatic Social Media Content Labeling and Classification based on BERT Neural Network

I-Hsien Ting
Department of Information Management
National University of Kaohsiung
Kaohsiung, Taiwan
iting@nuk.edu.tw

Chia Sung Yen
Department of Cultural and Creative Industries
National Chinyi University of Technology
Taichung, Taiwan
csyen@ncut.edu.tw

Chia-Chun Kang
Department of Comput. Sci. and Info. Engr.
Shu-Te University
Kaohsiung, Taiwan
kcc0211@stu.edu.tw

Shu-Chen Yang
Department of Information Management
National University of Kaohsiung
Kaohsiung, Taiwan
henryyang@nuk.edu.tw

***Abstract*—Web flow now is a very important success factor for social media marketing and thus more and more approaches for creating high web flow have been proposed in recent years. Automatic content generation (ACG) website is one of the possible approaches which can help to create web flow. In order to achieve the idea of automatic content generation website, web article classification has been considered the most important task. Therefore, we have development an empirical study to test the content labeling and article classification performance, which is based on the technique of BERT neural network. The performance evaluation including accuracy performance and time performance that are important for us to understand the possibility for implementing the ACG website in real environment, especially the possibility when dealing with large amount of data.**

***Index Terms*—Natural Language Processing, BERT, Neural Network, Social Media, Content Labelling, Content Classification**

I. Introduction

Website with the function of automatic content generation (ACG) has been claimed able to attract web flow as well as to reduce the human input for website management [1]. Once the flow oriented website has been created, it would be very useful for the purpose of online marketing. The most important factor for an ACG website is the classification of the content sources including News, content from social media and discussion forum, etc.

For content classification, BERT based neural network has been confirmed should be the most efficient approach with high performance at the current stage. In the process of BERT neural network

In order to have a success ACG website, we therefore would like to have an empirical study in this paper about the labeling and classification performance. In this paper, we will show an experiment in a real environment to demonstrate the process, the system for keyword labeling, neural network model training and testing as well as the performance evaluation.

The remainder of the paper is organized as follows. In section 2, related literature and works will be reviewed. The automatic web content generation and opinion system will be introduced in section 3. In section 4, we will introduce the design of the empirical study of automatic social media content labeling and classification based on BERT neural network. The results of the empirical study will be introduced in section 5 as well as the paper will be concluded in section 6.

IEEE/ACM ASONAM 2022, November 10-13, 2022
978-1-6654-5661-6/22/$31.00 © 2022 IEEE

II. Literature Review and Related Works

As discussed in the section of introduction, automatic content generation website is very important now for creating artificial web flow. In [1], the authors have proposed a system architecture about automatic content generation website. In the system, there are some important tasks, such as automatic keywords extraction, article classification, automatic content generation, website content formatting and presentation. Among these tasks, article classification is considered as the most important task to achieve the goal of automatic content generation website.

Document or Article classification is one of the application in the research area of Natural Language Processing, it is also called text mining in some researches. Article classification has been developed for many years and therefore there are many techniques can be used to achieve the idea. For example, data mining techniques, machine learning techniques, neural networks and deep learning techniques, etc [8], [9].

About keyword extraction, the details have been discussed in [2], [7]. For extracting the keywords, a sentence must be

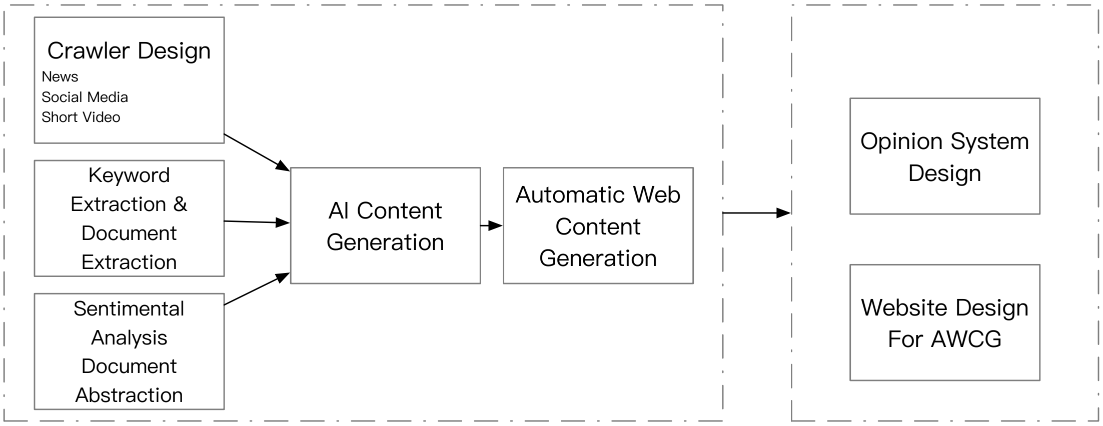

Crawler Design
News
Social Media
Short Video
Keyword Extraction & Document Extraction
Sentimental Analysis Document Abstraction
AI Content Generation
Automatic Web Content Generation
Opinion System Design
Website Design For AWCG

Fig. 1. The System Architecture of Automatic Web Content Generation and Opinion System

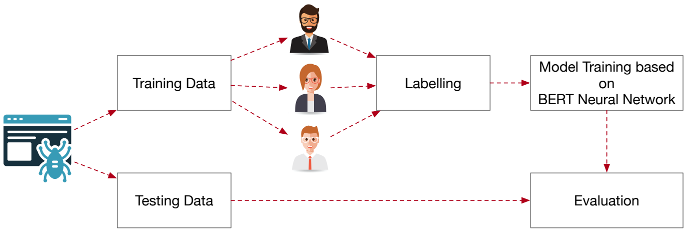


Fig. 2. The Process of the Empirical Study

For the training data, each article will be labeled manually by human input. Fro each article, at least 3 keywords will be labeled. After labeling process, BERT neural network model will be used to train the data. During the train phase, self-evaluation is performed continuously by using 10 folds cross validation. Finally, the empirical study will use the testing data to perform the final evaluation to test the performance of the classification by using BERT neural network model.

Figure 3 show the services and APIs deployment of the system that used for empirical study. From the architecture, the services are distributed in six databases and stored in 3 different databases.

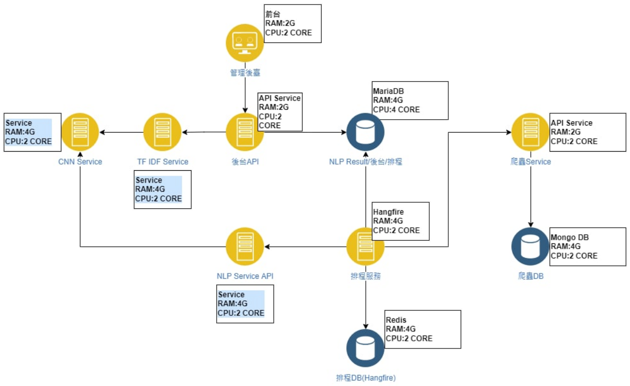
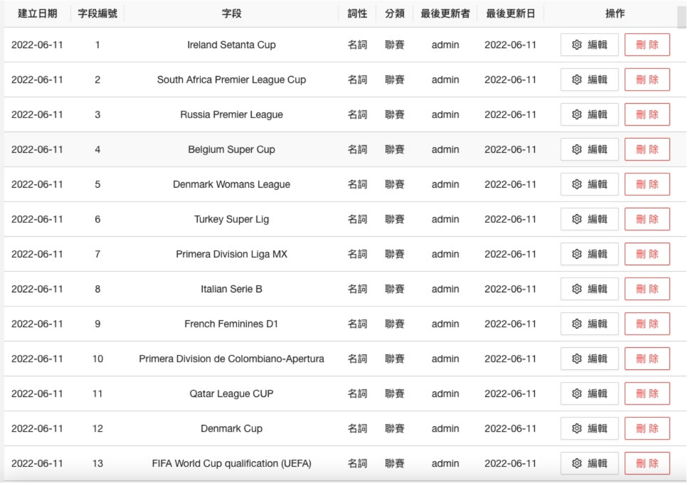


Fig. 3. The Services and APIs Deployment of the System

Figure 4 is the interface for keywords labeling. For each article, the user can edit the keywords or delete the labeled keywords.



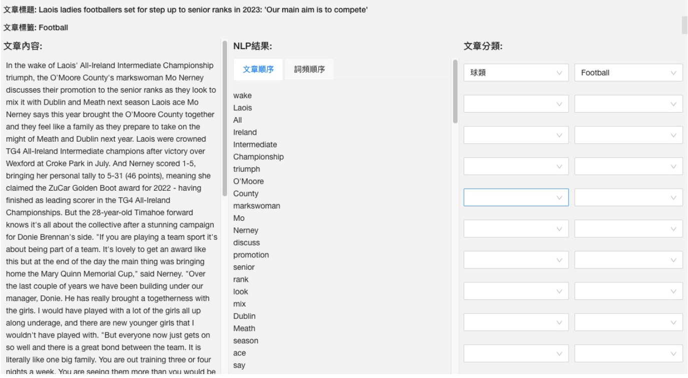
建立日期	字段編號	字段	詞性	分類	最後更新者	最後更新日	操作
2022-06-11	1	Ireland Setanta Cup	名詞	聯賽	admin	2022-06-11	⚙ 編輯 刪除
2022-06-11	2	South Africa Premier League Cup	名詞	聯賽	admin	2022-06-11	⚙ 編輯 刪除
2022-06-11	3	Russia Premier League	名詞	聯賽	admin	2022-06-11	⚙ 編輯 刪除
2022-06-11	4	Belgium Super Cup	名詞	聯賽	admin	2022-06-11	⚙ 編輯 刪除
2022-06-11	5	Denmark Womans League	名詞	聯賽	admin	2022-06-11	⚙ 編輯 刪除
2022-06-11	6	Turkey Super Lig	名詞	聯賽	admin	2022-06-11	⚙ 編輯 刪除
2022-06-11	7	Primera Division Liga MX	名詞	聯賽	admin	2022-06-11	⚙ 編輯 刪除
2022-06-11	8	Italian Serie B	名詞	聯賽	admin	2022-06-11	⚙ 編輯 刪除
2022-06-11	9	French Feminines D1	名詞	聯賽	admin	2022-06-11	⚙ 編輯 刪除
2022-06-11	10	Primera Division de Colombiano-Apertura	名詞	聯賽	admin	2022-06-11	⚙ 編輯 刪除
2022-06-11	11	Qatar League CUP	名詞	聯賽	admin	2022-06-11	⚙ 編輯 刪除
2022-06-11	12	Denmark Cup	名詞	聯賽	admin	2022-06-11	⚙ 編輯 刪除
2022-06-11	13	FIFA World Cup qualification (UEFA)	名詞	聯賽	admin	2022-06-11	⚙ 編輯 刪除

Fig. 4. The Interface of Keywords labelling System

Figure 5 is the interface of article classification and correction system. From figure 5, the details of keywords labelling for each article is shown. In the right side of the interface is the content of the article and the labeled keywords is in the middle part of the interface. In the left side, it shows the result of classification by using BERT neural network. The system administration can also edit the classification to made correction to the results for re-training.

V. The Results of the Empirical Study

In this section, we will show the results of the empirical study about the accuracy performance and time performance.



文章標題: Laois ladies footballers set for step up to senior ranks in 2023: 'Our main aim is to compete'

文章標籤: Football

文章內容:

In the wake of Laois' All-Ireland Intermediate Championship triumph, the O'Moore County's markswoman Mo Nerney discusses their promotion to the senior ranks as they look to mix it with Dublin and Meath next season Laois ace Mo Nerney says this year brought the O'Moore County together and they feel like a family as they prepare to take on the might of Meath and Dublin next year. Laois were crowned TG4 All-Ireland Intermediate champions after victory over Wexford at Croke Park in July. And Nerney scored 1-5, bringing her personal tally to 5-31 (46 points), meaning she claimed the ZuCar Golden Boot award for 2022 - having finished as leading scorer in the TG4 All-Ireland Championships. But the 28-year-old Timahoe forward knows it's all about the collective after a stunning campaign for Donie Brennan's side. "If you are playing a team sport it's about being part of a team. It's lovely to get an award like this but at the end of the day the main thing was bringing home the Mary Quinn Memorial Cup," said Nerney. "Over the last couple of years we have been building under our manager, Donie. He has really brought a togetherness with the girls. I would have played with a lot of the girls all up along underage, and there are new younger girls that I wouldn't have played with. "But everyone now just gets on so well and there is a great bond between the team. It is literally like one big family. You are out training three or four nights a week. You are seeing them more than you would be

NLP結果:

文章順序 詞頻順序

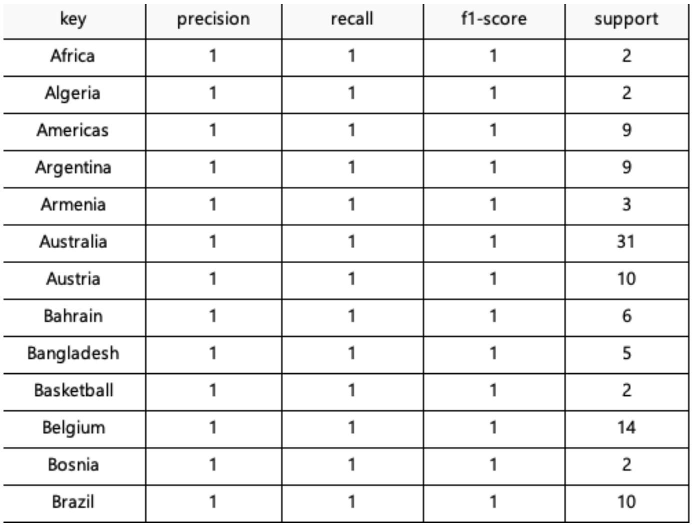
wake
Laois
All
Ireland
Intermediate
Championship
triumph
O'Moore
County
markswoman
Mo
Nerney
discuss
promotion
senior
rank
look
mix
Dublin
Meath
season
ace
say

文章分類:

球類 Football

Fig. 5. The Interface of Article Classification and Correlation System

Figure 6 is the performance of keyword extraction performance. It is used for the training phase to check the accuracy of the keyword that extracted by system and the keyword that labeled by users. The field ”key” in figure 6 means the keyword and precision, recall and F1-score are measurements for measuring the accuracy performance. In the figure, the field ”support” means how many times the keyword extracted from the entire articles for training. We will continue to adjust the labeled keywords until the accuracy performance reach a reasonable threshold, which is average 0.9 for F1-score.



key	precision	recall	f1-score	support
Africa	1	1	1	2
Algeria	1	1	1	2
Americas	1	1	1	9
Argentina	1	1	1	9
Armenia	1	1	1	3
Australia	1	1	1	31
Austria	1	1	1	10
Bahrain	1	1	1	6
Bangladesh	1	1	1	5
Basketball	1	1	1	2
Belgium	1	1	1	14
Bosnia	1	1	1	2
Brazil	1	1	1	10

Fig. 6. The Performance of Keyword Extraction Performance

Table 1 shows the accuracy performance evaluation of classification and table 2 shows the time performance evaluation of classification.

In table 1, the value are average vale of entire articles. It show the precision, recall and F-1 score are all reach 0.99, which means the accuracy is very high. The value of average number of articles means the average number of articles for each category.

In table 2, we are trying to test the time performance under different computer specifications, different number of labeled keywords by fixing the number of articles. For computer

TABLE I
CLASSIFICATION ACCURACY PERFORMANCE EVALUATION

Measurement	Value
Precision	0.9993742
Recall	0.9957561
F-1 Score	0.9973121
Average Number of Articles	231.24725

TABLE II
CLASSIFICATION TIME PERFORMANCE EVALUATION

Computer Specifications	Articles	Labels	Time (second)
GCP-GPU (8 CPU/60G RAM/c100 16G GPU)	12086	1000	2100
GCP-GPU (8 CPU/60G RAM/c100 16G GPU)	12086	3000	4430
GCP-GPU (32 CPU/64G RAM)	12086	1000	700
GCP-GPU (32 CPU/64G RAM)	12086	3000	1334
DEV (32 CPU/64G RAM)	12086	1000	983
DEV (32 CPU/64G RAM)	12086	3000	2305

specifications, GCP means Google Cloud Platform (Computer Engine Service) and DEV means desktop computer. From the results, it shows the time performance is better when performing the process in cloud environment.

VI. CONCLUSION

In this paper, we have design an empirical study of automatic social media content labeling and classification based on BERT neural network model. In the empirical study, the performance of keyword labelling and article classification have been evaluated, including accuracy performance and time performance. From the results of the empirical study, we firstly proved that the designed system architecture of automatic web content generation and opinion system is actionable in real environment. Due to article classification is the critical take of the system. Secondly, we also show the performance of accuracy and time are both very high. Finally, we also show the time performance evaluation, which would be very useful for whom want to take the system in practice to select suitable computer specifications.

The direction of future research may consider to evaluate the performance of web flow when performing the system in real environment to build automatic content generated website. Researchers may also want to focus on different category of articles, due to the articles are limited to sport related in this paper.

REFERENCES

[1] I. -H. Ting, C. -S. Yen, "Towards Automatic Content Generated Website Based on Content Classification and Auto-article Generation" In Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '21). Association for Computing Machinery, New York, NY, USA, 436–438. https://doi.org/10.1145/3487351.3488414.

[2] I. -H. Ting, S. -C. Yang, C. -S. Yen and T. -H. Tsai, "Hot Topics Detection by Using 2-Layers Keywords Extraction" 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2020, pp. 926-928, doi: 10.1109/ASONAM49781.2020.9381308.

[3] A. Adhikari, A. Ram, R. Tang, J. Lin, "Docbert: Bert for document classification" arXiv preprint arXiv:1904.08398, 2019

[4] C. He, S. Chen, S. Huang, J. Zhang and X. Song, "Using Convolutional Neural Network with BERT for Intent Determination," 2019 International Conference on Asian Language Processing (IALP), 2019, pp. 65-70, doi: 10.1109/IALP48816.2019.9037668.

[5] J. Devlin, M. -W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". NAACL-HLT (1) 2019: 4171-4186

[6] T. Huang, Q. She, J. Zhang, "BoostingBERT: Integrating Multi-Class Boosting into BERT for NLP Tasks" CoRR abs/2009.05959 (2020)

[7] W. Jin, H. H. Ho, R. K. Srihari. 2009. "OpinionMiner: a novel machine learning system for web opinion mining and extraction." In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09). Association for Computing Machinery, New York, NY, USA, 1195–1204. https://doi.org/10.1145/1557019.1557148

[8] M. Zeppelzauer, D. Schopfhauser, "Multimodal classification of events in social media", Image and Vision Computing, Volume 53, 2016, Pages 45-56, ISSN 0262-8856

[9] S. Kinsella, A. Passant, J. G. Breslin, "Topic Classification in Social Media Using Metadata from Hyperlinked Objects". Advances in Information Retrieval. ECIR 2011. Lecture Notes in Computer Science, vol 6611. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-20161-520

[10] S. Gupta, S. E. Bolden, J. Kachhadia, A. Korsunska, J. Stromer-Galley, "PoliBERT: Classifying political social media messages with BERT." Paper presented at the Social, Cultural and Behavioral Modeling (SBP-BRIMS 2020) conference. Washington, DC, October 18-21, 2020.