Automatic Detection of Fake Crowdfunding Projects

Qi Li[1] and Jian Qu[2]

[1,2]Faculty of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi, Thailand
E-mail: 6372100118@stu.pim.ac.th, jianqu@pim.ac.th

Received: April 28, 2022 / Revised: June 10, 2022 / Accepted: October 18, 2022

***Abstract*—There may be fake information in some crowdfunding projects. However, it is difficult for crowdfunding platforms and investors to find fake information in crowdfunding projects. At present, many scholars have studied the methods for identifying fake information, but most of them studied how to distinguish fake information from news articles. Therefore, this research focuses on how to identify fake information that may exist in crowdfunding projects. The detection of fake crowdfunding projects includes functions such as keyword extraction, external knowledge extraction, and classification of real and fake projects. To identify possible fake information in the crowdfunding project, we need to understand more about the crowdfunding project by extracting the keywords of the crowdfunding projects. Therefore, this research compared TF-IDF, CKPE, YAKE, RAKE, TextRank4zh, FastTextRank, HarvestText, and BERT pre-training model methods. We used precision, recall, and F1 scores to measure the effectiveness of the keyword extraction method. Then, we obtained features for judging the authenticity of crowdfunding projects by extracting external knowledge of keywords. Finally, projects were classified using a classification algorithm. The validity of this study for the classification of fake crowdfunding projects achieves 83.77% by the NB method in the dataset.**

***Index Terms*—BERT Model, Crowdfunding Project, Information Extraction, Keywords Extraction, Web-Data**

I. Introduction

Today, crowdfunding projects are rising in fervor. As crowdfunding projects come into the public eye, some people will use crowdfunding projects to scam the funds raised for the project, hence the rise in the number of fake crowdfunding projects. According to statistics, the rate of project failures and fraud among investors exceeds 64.6% [1]. After the study on fraudulent crowdfunding projects, we divided fraudulent crowdfunding projects into two types. The first type is that the projects are theoretically and logically achievable, but the initiator of the project maliciously defrauds the crowdfunding funds. This type of fraudulent project is due to the reputation of the project's initiators themselves. The second type is the fake description and design of the crowdfunding project by the initiator of the project. This type of fraudulent project has contradictions and loopholes in the logic of the implementation, which cannot be achieved with the currently available technology. For example, the famous fake project-Triton underwater respirator[1]. The initiator of the project claimed to be able to make Triton, a small artificial fish gill. This respirator produces oxygen from water by electrification. Furthermore, the size is small, and someone can carry it. The developers of Triton claimed their device would allow a user to breathe underwater for 45 minutes at a maximum depth of 15 feet. However, the project is a copy of an idea from a science fiction movie, and Dr. Alistair Dove has confirmed in Deep Sea News that the project is impossible to achieve[2].

On the one hand, it is difficult for the average investor to identify fraudulent information in crowdfunding projects because fraudulent crowdfunding projects are usually deliberately falsified by fraudsters. Another hand, it is difficult for investors with some relevant professional knowledge to identify fraudulent information for crowdfunding projects because crowdfunding projects are usually novel and creative. For example, most of Triton's investors are divers with professional knowledge, but they were still defrauded. Thus, identifying fraudulent crowdfunding projects is a challenging task. Therefore, few people are devoted to the detection of fraudulent crowdfunding projects, and most scholars are devoted to the detection of fake news.

During the research, we found that some ideas from the method of fake news detection can be applied

[1] https://gearjunkie.com/news/triton-artificial-gills-breathe-underwater
[2] https://www.deepseanews.com/2014/01/triton-not-dive-or-dive-not-there-is-no-triton/

to the research of detecting fraudulent crowdfunding projects. For example, a method of fake news detection based on article information. Text information extraction can help us learn more information about crowdfunding projects, which can help us further research the detection of fake crowdfunding projects. The important information in the text can generally be reflected in the keywords. Therefore, we tried to obtain the feature information of crowdfunding projects by extracting the keywords of crowdfunding projects, and this research also compared different keyword extraction methods. However, extracting crowdfunding project information alone cannot help us identify fraudulent crowdfunding projects, so we proposed a comprehensive method to identify fraudulent crowdfunding projects.

The method proposed by the research first extracted the keywords of the crowdfunding project and then retrieved the characteristic information of the crowdfunding project to obtain the judgment basis of the fraudulent crowdfunding project. Finally, we used machine learning algorithms to classify the authenticity of crowdfunding projects.

II. Guidelines for Manuscript Preparation

The methods of keyword extraction generally include unsupervised and supervised methods.

A. Unsupervised Method

The unsupervised method extracted keywords without manual annotation of the corpus. The unsupervised method includes two approaches. (1) The first approach is the method of extracting keywords based on statistical features. Such methods mainly include TF-IDF [2], YAKE [3], and other methods. These methods do not require training data, and mainly use the position of words in the document, co-occurrence frequency [4], Term Frequency (TF), Inverse Document Frequency (IDF), N-gram, and PAT tree as statistical features to select terms as keywords [5]. This type of method can exclude words that are not relevant to the text and is fast to implement, but it cannot reflect the lexical organization structure within the article. Therefore, Qu et al. [6]. proposed to use symmetric conditional probability, chi-square, correlation measure, number of segments, and distance as statistical features to select terms as keywords, reflecting the relationship between words within the article [6]. (2) The second approach is the method that was based on graph networks. These methods built a graph network based on words or phrases. Then these methods used algorithms to calculate important nodes as keywords. Among them, widely used methods are the graph-based ranking model (TextRank) [7] and the rapid keyword extraction algorithm (RAKE) [8]. TextRank algorithm is an important ranking algorithm that can extract keywords, keyphrases, and key sentences from documents. This method inherited the idea of PageRank. Compared with the method of TF-IDF, it can make full use of the relationship between text elements. However, this method still does not solve the problem that high-frequency words excessively affect the results.

In contrast, the RAKE algorithm introduced a concept of degree, did not make any distinction between words and phrases, and used the co-occurrence information of words to determine keywords. The algorithm used Word Frequency, Word Degree, and the ratio of the degree to frequency as features for keyword extraction. However, this method still has a strong dependence on the list of deactivated words. Since the unsupervised keyword extraction method cannot synthesize multiple information of the text in text information extraction, it is not very effective for keyword extraction.

B. Supervised Method

To improve the effectiveness of keyword extraction, supervised keyword extraction methods are proposed. Supervised learning uses the model learned from a set of trained texts to extract keywords. The supervised method has two aspects. The first type is based on traditional supervised learning methods, which include Naive Bayes (NB), Decision Tree (DT), Logistic Regression (LR), K Nearest Neighbors (k-NN), Support Vector Machine (SVM) [9], and methods based on Learning to Rank (LTR). Zhang et al. [9] have proposed an SVM-based method for keyword extraction, which uses “global context information” and “local context information” to extract keywords from documents. However, these methods have certain limitations in adaptability.

To enhance the adaptability of the method, Qu et al. proposed a method that combines statistical features and adaptive rules to extract keywords [10]. The adaptive rule was to modify the basic rule through each term item based on the basic rule (the predefined regular expression matching rules) and then form an adaptive regular expression rule. Another study by Qu [11] proposed that when the search distance was eight, the key characters presented the best recall rate. Thus, a rule-based algorithm that automatically adjusts the candidate generation system can effectively improve the efficiency of candidate selection.

However, the keyword extraction algorithm combining statistical features and adaptive rules relies more on statistical features, and its rules have limitations when applied to other types of text. Therefore, the second type of keyword extraction based on deep learning was proposed. In 2016, Zhang et al. proposed a new deep Recurrent Neural Network (RNN) model, which can jointly handle keyword ranking and keyphrases generation tasks [12]. However, this model is limited in its perf Meng, Zhao et al. proposed a keyphrases generation model

(CopyRNN) with an encoder-decoder framework [13]. This model not only captured the semantic meaning behind the text but also generated missing key phrases based on the text semantics.

However, the RNN method has a strong dependence on the calculation of the sequence in the training process. The distance between time steps may lead to the problem of gradient disappearance, which can be well solved by the BERT model. This model was a typical two-way coding model. It used Transformers as the main framework of the algorithm to capture the two-way relationship in sentences, which improved the model's language representation ability and feature extraction ability [14]. The BERT model has been open-sourced tool by Google, and researchers can use the BERT pre-training model for their natural language processing research. Such as paraphrase recognition, semantic text similarity, repeated question detection, and question-answer retrieval. The BERT model provided a simple model (BERT-base), a complex model (BERT-large), and many embedded models.

This study applied the BERT pre-training model to simplified Chinese text processing and compared the effects of different BERT pre-training models on keyword extraction.

The main methods for the detection of fake crowdfunding projects are divided into two categories: First, the language approach, language patterns linked to false (contradictory); P E Rez -Rosas, who proposed language method to detect conflicts. Second, the network method, using network information to fake (contradictory) connections. Language methods include (1) Data representation methods. It uses the bag-of-words method to aggregate and analyzes the frequency of words or multiple words to reveal contradictions. However, this method not only relies on the language but also the isolated n-gram, which will be out of touch with the contextual information. (2) In-depth syntax method. Various studies have shown that the analysis of words alone is not enough to predict contradictions. In-depth grammatical analysis can predict contradictions. The in-depth grammatical analysis is realized by structural probability context-free grammar, which converts sentences into rewriting rules to describe the grammatical structure of sentences. The accuracy of this method to detect contradictions is 85%-91%, and the specific accuracy depends on the type of rule used. Besides, you can also rely on third-party tools for in-depth syntax analysis to achieve contradiction detection, such as AutoSlog-TS syntax analyzer and other tools. However, this method alone is not enough to detect contradictions, and it needs to be combined with other language methods or network analysis techniques. (3) Discourse analysis, as an alternative method of contradictory clues, analyzes the contradiction through the degree of compatibility between information and information. The accuracy of this method for contradiction detection is 91%, but this method is limited to the application field 4. Rhetorical structure and discourse analysis. Discourse description is realized through the analysis framework of rhetorical structure theory, which identifies examples of rhetorical relations between language elements. But this method is too inferior in the accuracy of contradiction detection. Both of these methods are combined with machine learning to train classifiers to monitor and analyze. Hai et al. proposed semi-supervised learning that is a combination of language methods or network methods and machine learning methods.

III. Methods

This research compared several keywords extraction methods, such as TF-IDF, YAKE, RAKE, TextRank4zh, FastTextRank, HarvestText, and BERT pre-training models for a crowdfunding project. In addition, we tested five BERT pre-training models. See Table VIII for details. Furthermore, we performed external information retrieval on the extracted keywords, and the retrieved information is used as a feature for judging fraudulent crowdfunding information. The features are trained by using traditional machine learning to identify fraudulent crowdfunding projects. The flowchart of this research is shown in Fig. 1.

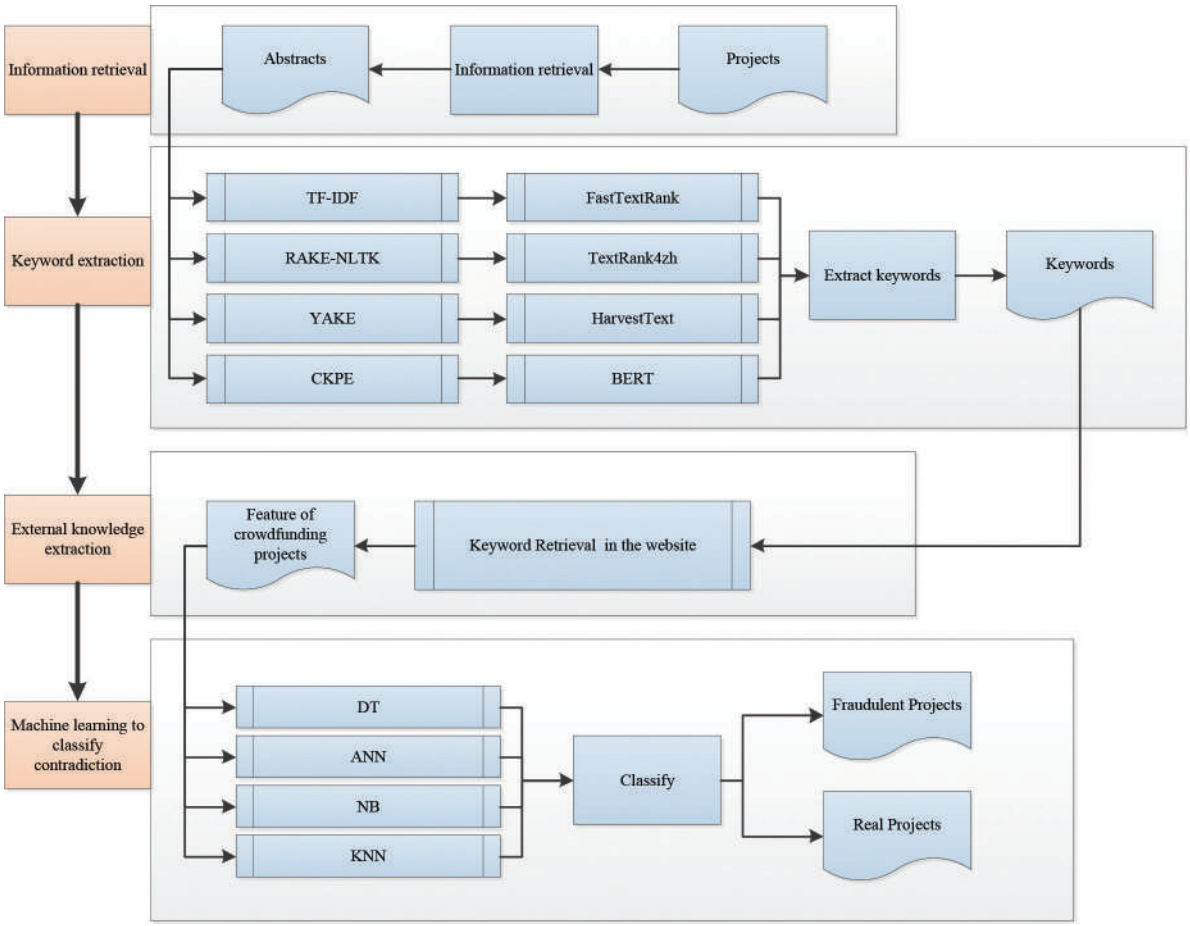

Fig. 1. The flowchart of the research

This study included four experimental steps, information retrieval, keyword extraction, external knowledge extraction, and machine learning for contradiction detection and classification. The specific steps are shown in Fig. 1. Among them, the information retrieval process is the project summary information feedback. In the process of keyword extraction, we compared the traditional statistical method and the deep learning keyword extraction method. The best results of keyword extraction were used for our experiments. We use keyword extraction for network knowledge extraction to obtain the features of crowdfunding projects to judge their authenticity of crowdfunding projects. In the final step, we used the classification algorithms to classify the features of crowdfunding projects obtained by Web-knowledge extraction to predict whether the projects are fake or not.

A. Keyword Extraction

1) TF-IDF

Initially, the TF-IDF algorithm was proposed to calculate the importance of words to documents. According to this algorithmic idea, if the weight of the word was higher when calculating the weight of feature words, the frequency of the word in the text is higher, the more likely it was a keyword [2]. Therefore, after calculating the weights of feature words, the weight values of all words are sorted. Then the method selects the Top K keywords with the largest weights or the keywords with weights greater than a certain threshold. Table I is an example of using TF-IDF to extract the keywords.

TABLE I
EXAMPLE PF EXTRACTING KEYWORDS BY TF-IDF ALGORITHM

Construct	Data
Source Input	MATE X是世界上最酷的可折叠电动自行车。在世界自行车之都丹麦哥本哈根，人们只想踩踏板就对它进行了构想和设计……升级后的制动更加平稳，所需制动力更少。 ***Translation:*** **"**MATE X is the coolest foldable electric bike in the world. It was conceived and designed in Copenhagen, Denmark, the cycling capital of the world, where people just wanted to pedal …… the upgraded brakes are smoother and require less stopping power."
Output KeyWords	[‘制动’，‘可折叠’，‘自行车’，‘升级’ ‘踏板’，‘显示器’，‘动力’，‘电动’，‘功率’，‘配备’] ***Translation:*** [‘brake’, ‘foldable’, ‘bike’, ‘upgrade’, ‘pedal’, ‘display’, ‘power’, ‘electric’, ‘power’, ‘equipped’]

Initially, the TF-IDF algorithm was proposed to calculate the importance of words to documents. According to this algorithmic idea, if the weight of the word was higher when calculating the weight of feature words, the frequency of the word in the text is higher, the more likely it was a keyword [2]. Therefore, after calculating the weights of feature words, the weight values of all words are sorted. Then the method selects the Top K keywords with the largest weights or the keywords with weights greater than a certain threshold. Table I is an example of using TF-IDF to extract the keywords.

2) YAKE

YAKE is an unsupervised method of automatically extracting keywords based on text features [3]. The keywords extractor, which relies on the statistical characteristics of the text in a single document, selects the most important keywords in the text. This method does not need to be trained on a specific set of text documents, nor does it rely on dictionaries and external corpora. Based on those advantages, we used YAKE for keywords extraction of crowdfunding projects.

YAKE Keywords extractor has been currently used as a keyword's extraction tool. YAKE can be used as a CLI utility in a Docker container or as a REST API server in a Docker container [3]. However, this study used YAKE as a standalone tool and used YAKE as a keyword's extraction tool in python. This method also provided an end-to-end keyphrases extraction pipeline to extract keywords from text documents.

YAKE Keywords extractor can extract keywords for various languages. As shown in Table II, its input is a source input in the example of Table I. The keyphrases column (Table II) is the keyphrases extracted with the YAKE method. The score column (Table II) is the similarity score between the keywords/keyphrases and the text. If the score was lower, the keywords would be more relevant.

TABLE II
EXAMPLE PF EXTRACTING KEYWORDS BY YAKE ALGORITHM

Id	Keyphrases	The Similarity Score
1	多个支持者交付了 mate bikes ***Translation***：Multiple supporters delivered mate bikes	0.0226245446
2	mate bike 在成功的基础上, 我们将所有的时间, 精力, 血液和汗水投入到电动自行车的美丽之中, 这对于任何骑手, 任何旅程, 在任何特定情况下都是完美的 ***Translation***：Based on the success of mate bike, we put all our time, energy, blood, and sweat into the beauty of electric bikes that are perfect for any rider, any journey, and in any given situation.	0.0226245446
…	….	…
9	mate x 拥有功率高达 ***Translation***：mate x has power up to	0.1598191003
10	usd （价值） ***Translation***：usd (value)	0.1652103123

3) RAKE-NLTK

NLTK is a toolkit of natural language processing, which can greatly contribute to the field of natural language processing [15]. The toolkit is very powerful, with powerful functions for text processing, such as text classification, text tagging, stemming, semantic inference, and other functions.

TABLE III
EXTRACT KEYWORDS/KEYPHRASES BY RAKE-NLTK

ID	Keyphrases
1	我们还提供了仅需99 usd （ 价值149 usd ） 即可升级到彩色smart美丽显示器 （ 如图 ） 的选项 ***Translation***：We also have the option to upgrade to a colour smart beautiful display (pictured) for just 99 usd (worth 149 usd).
2	并在所有型号中均 配备了 令人印象深刻的48v电池和控制器 ***Translation:*** And comes with an impressive 48v battery and controller in all models.
…	…
9	该齿轮箱具有改进的易换档性能 ***Translation:*** The gearbox has improved easy shifting performance.
10	该新的计算机显示器配备有背光led作为标准 ***Translation***：This new computer monitor is equipped with backlit led as standard.

In this research, we combined the rapid automatic keywords (RAKE) [8] algorithm with the NLTK toolkit [16]. It formed a powerful keyword extraction method, which was called RAKE-NLTK. As shown in Table III, the keywords/keyphrases are extracted by using this method. Its input is a source input in example Table I.

4) TextRank4zh

The TextRank4zh method is a keyword extraction method [7]. The method first splits the original text into sentences, then filters the stop words of the sentence, and last retains the words of the specified part of speech [7]. The method was to calculate the importance of word nodes according to the principle of the graph to obtain keywords and key phrases.

An example of using the TextRank4zh method to extract keywords is shown in Table IV. The input for this example is the source input of Table I. The weight scores of keywords are listed in the score column (Table IV). According to the TextRank4zh, the higher the weight score, the more important the keywords.

TABLE IV
EXTRACT KEYWORDS BY TextRank4zh AND THE SCORE OF KEYWORDS

Id	Keywords	Score
1	Mate	0.028890332978898956
2	功率 ***Translation***：*p*ower	0.024104175996508638
3	…	…
9	电动 ***Translation***：*e*lectric	0.01716594583001619
10	构想 ***Translation***： conception	0.016248380651252663

5) FastTextRank

FastTextRank is divided into FastTextRank Word and FastTextRank Sentence. FastTaxtRank Word is a method that divides the article into sentences. It calculates the similarity between words, constructs graphics according to the word similarity, and calculates the importance of each word by an iterative algorithm to obtain keywords [16]. This method is based on the TextRank graphics algorithm. FastTaxtRank Sentence can extract sentences. The method calculates the similarity between sentences by the cosine similarity of word vectors.

TABLE V
EXTRACT KEYWORDS AND KEYPHRASES BY FASTTEXTRANK

Id	Keywords	Keyphrases
1	到 ***Translation:*** reach	我们还升级了Tektro制动系统，包括杠杆和 卡钳，以实现更快，更有效的制动力。 ***Translation:*** We also upgraded the Tektro braking system, including levers and calipers for faster, more effective stopping power.
2	功率 ***Translation:*** power	新款MATE X拥有功率高达750W的强大动力，并在所有型号中均 配备了 令人印象深刻的48V电池和控制器。 ***Translation:*** The new MATEX is packed with power up to 750W and comes with an impressive 48V battery and controller in all models.
…	…	…
9	配备 ***Translation:*** equipped	可折叠踏板是一个全新的设计封装在一个更坚固的踏板架为改进的功率转移。 ***Translation:*** The foldable pedals are an all-new design packaged in a sturdier pedal rack for improved power transfer.
10	USD	在世界自行车之都丹麦哥本哈根，人们只想踩踏板就对它进行了构想和设计。 ***Translation:*** It was conceived and designed in Copenhagen, Denmark, the cycling capital of the world, where people just wanted to pedal.

The word vectors are obtained from each sentence that was being compared. Table V is an example of using FastTextRank to extract keywords and keyphrases, and its input is the source input in Table I.

6) HarvestText

HarvestText is an unsupervised method. The analysis of specific domain text can be processed by the HarvestText method that can integrate domain knowledge (types, aliases) [17]. The HarvestText method extracts keywords based on the TextRank algorithm. The HarvestText method can also use dependency grammar (DG)[1] to extract semantic triple[2] that may represent events.

[1] https://en.wikipedia.org/wiki/Dependency_grammar
[2] https://en.wikipedia.org/wiki/Semantic_triple

TABLE VI
SYNTACTIC STRUCTURE ANALYSIS RESULTS BY HARVESTTEXT MEHTOD

Id	Word literal value/entity name	Part-of-speech	Dependency	Dependent sub-words
0	MATE 是 (MATE is)	v	主谓关系 (subject-predicate relationship)	6
1	世界(world)	n	定中关系 (Ding-China relations)	2
2	上(on)	f’	状中结构 (mesostructure)	6
3	‘酷(cool)	a	主谓关系 (subject-predicate relationship)	6
4	的	u	右附加关系 (right append relationship)	3
5	可(can)	v	状中结构 (mesostructure)	6
6	折叠 (foldable)	v	核心关系 (core relationship)	-1
7	电动 (electric)	b	定中关系 (Ding-China relations)	8
8	自行车 (bike)	o	动宾关系 (verb-object relationship)	6
9	。	w	标点符号 (Punctuation)	6

We analyzed the dependency grammar of crowdfunding projects by the HarvestText method and extracted semantic triple to represent events. The method obtains the key sentences of the crowdfunding projects by combining the triples. To extract facts from sentences is based on dependency grammar. Firstly, the method found meaningful triples in the sentence using a subject, predicate, and other syntactic relationships. Then Semantic triple, which centered on the predicate, were extracted. The extracted sentences that may represent the event included three syntactic structures: Subject-Verb-Object (SVO) structure, Object–Verb–Subject (OVS) structure, and Subject–Object–Verb (SOV) structure. The stop word list in this method is Baidu stop words by default. Syntactic structure analysis results are shown in Table VI.

Table VI is an example of fact extraction using the HarvestText method based on dependency grammar. This example performed part-of-speech tagging and dependency analysis on words in a sentence and then extracted semantic triple that may express events. For example, the original sentence: “MATE X 是世界上最酷的可折叠电动自行车。 (MATEX is the coolest foldable electric bike in the world.)” The proposed semantic triple is: (‘MATE X 是’, ‘折叠’, ‘电动自行车’), (‘MATE X is’, ‘foldable’, ‘electric bike’) as shown in Table VII.

TABLE VII
EXTRACT SEMANTIC TRIPLE BY HARVEST TEXT METHOD

Construct	Data
Input sentence	MATE X 是世界上最酷的可折叠电动自行车。 ***Translation:*** MATEX is the coolest foldable electric bike in the world.
Syntactic relation	[0, ’MATE X是’, ’v’, ’主谓关系’, 6] [1, ’世界’, ’n’, ’定中关系’, 2] [2, ’上’, ’f’, ’状中结构’, 6] [3, ’最酷’, ’a’, ’主谓关系’, 6] [4, ’的’, ’u’, ’右附加关系’, 3] [5, ’可’, ’v’, ’状中结构’, 6] [6, ’折叠’, ’v’, ’核心关系’, -1] [7, ’电动’, ’b’, ’定中关系’, 8] [8, ’自行车’, ’设备’, ’动宾关系’, 6] [9, ’。’, ’w’, ’标点符号’, 6]
Event triplet	MATE X 是折叠电动自行车 ***Translation:*** MATE X is a folding electric bike.

7) BERT

BERT (Bidirectional Encoder Representations from Transformers) is a pre-training technology for natural language processing proposed by Google. BERT is a deep two-way, the unsupervised model that only uses a plain text corpus for training [18].

This research used the BERT pre-training model to extract keywords from the text. In extracting keywords, firstly, BERT used the Count Vectorizer in Scikit-Learns to remove stop words and extract keywords/keyphrases candidates [18]. Secondly, it converted documents and candidates into vectors by embedding the BERT pre-trained model. Then it calculated the cosine similarity between the candidate vector and the document vector. Finally, keywords/keyphrases were selected according to the cosine similarity [18]. In this paper, the flow chart of applying the BERT pre-training model to extract keywords is shown in Fig. 2.

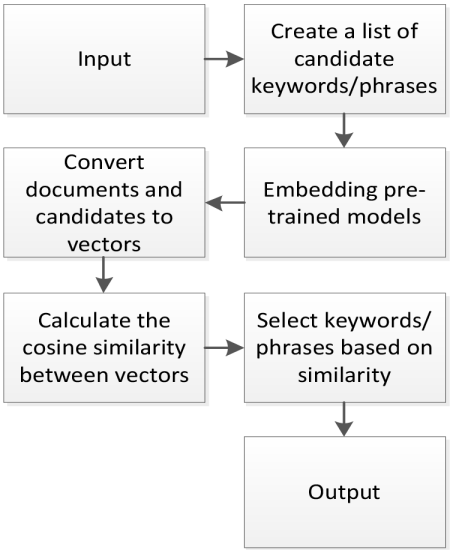

Fig. 2. Flow chart of extracting keywords with BERT pre-training model

When using the BERT pre-training model for keywords extraction, it did not need to train the model and only fine-tune the model to obtain the keywords extracted using different sentence transformer models. When using the CountVectorizer in Scikit-Learns to create a list of keywords/keyphrases candidates, we can customize the size of the candidate words and the list of stop words. When calculating cosine similarity, we can also use different algorithms (maximum and similarity, maximum marginal correlation) to diversify the results [18].

TABLE VIII
BERT PRE-TRAINING MODEL

Label	Model	Annotation
BERT-1	*Distilbert-base-nli-mean-tokens*	Semantic similarity model
BERT-2	*Distiluse-base-multilingual-cased-v1*	Multilingual knowledge distilled version 1 of multilingual Universal Sentence Encoder
BERT-3	*Distiluse-base-multilingual-cased-v2*	Multilingual knowledge distilled version 2 of multilingual Universal Sentence Encoder
BERT-4	*Paraphrase-distilroberta-base-v1*	Paraphrase recognition model
BERT-5	*Paraphrase-xlm-r-multilingual-v1*	Multilingual version of paraphrase-distilroberta-base-v1

Many BERT pre-training models have been proposed, and each model was good in different fields. This study analyzed and compared five models in the multi-language general model. In this study, we shorten the BERT model names from BERT-*1* to BERT-*N*. The information on each BERT pre-training model is shown in Table VIII.

B. External Web-Knowledge Extraction

This research proposed a method to extract the features of crowdfunding projects. The method used the feedback results of retrieving external knowledge as features of whether the crowdfunding projects are fake or not. The feature items we retrieved are listed in Table IX. The Web-knowledge retrieval was done by obtaining external knowledge related to the project keywords on the web [19]. Keywords represent some features of an item, retrieving external knowledge about keywords can help us learn more information about whether the technology of the project can be achieved or not. Since, the initiator of crowdfunding wanted to obtain financial support for the realization of the project through crowdfunding. If the current technology cannot realize the project, the project is more likely to be a fraudulent project. Therefore, we retrieved the keywords that can represent the characteristic information of the crowdfunding project to determine whether the current technology can realize a certain characteristic function of the project.

TABLE IX
FEATURE WE RETRIEVED

Feature-ID	Feature	Example	Translation
1	Keyword and "Category"	可折叠 and "自行车"	Foldable and "bike"
2	"Keyword" and "Category"	"可折叠" and "自行车"	"Foldable" and "bike"
3	"Keyword" and Category	"可折叠" and 自行车	"Foldable" and bike
4	Keyword and Category	可折叠 and 自行车	Foldable and bike
5	"Keyword"	"可折叠"	"Foldable"
6	Keyword	可折叠	Foldable

C. Machine Learning

This research used the results of retrieving external knowledge feedback as features to select and classify projects using traditional machine learning. In this research, the effectiveness of classification with Decision Tree (DT), Naïve Bayes (NB), Artificial Neural Network (ANN), Support Vector Machines (SVM) and other algorithms were compared. The specific results are shown in section IV.

IV. RESEARCH METHODS EXPERIMENT RESULT AND DISCUSSION

A. Data Set

There are two data sets for this research. Data set one contained 120 crowdfunding projects from a list of the highest-funded crowdfunding projects in Wikipedia[1]. For better comparison with dataset two, we only selected the project from Kickstarter and Indiegogo. There are also no restrictions on the types of crowdfunding projects. Data set two contained 100 crowdfunding projects, which came from Kickstarter's and Indiegogo's official webpage. For this dataset, we selected 20 fake crowdfunding projects and 80 real crowdfunding projects. We collected a hundred datasets based on the popularity of crowdfunding on Kickstarter and Indiegogo. Only 20 fake crowdfunding project datasets are collected since fake crowdfunding projects are difficult to found. In this study, a fake crowdfunding project refers to a project similar to Triton where the promoters falsely describe and design the project, but there are logical contradictions and loopholes.

[1] https://en.wikipedia.org/wiki/List_of_highest-funded_crowdfunding_projects

B. Results of Keywords Extraction

This research used TF-IDF, YAKE, RAKE-NLTK, TextRank4zh, FastTextRank, HarvestText, and BERT to extract keywords. According to the precision, recall, and F-score of different methods in extracting keywords to compare those method's effectiveness. The calculation formulas of precision, recall and F-score are shown in formulas (1), (2), and (3).

$$precison = \frac{\left|\{rel\ doc\} \cap \{ret\ doc\}\right|}{\{ret\ doc\}} \quad (1)$$

Precision is the ratio of relevant documents to retrieved documents.

$$recall = \frac{\left|\{rel\ doc\} \cap \{ret\ doc\}\right|}{\{rel\ doc\}} \quad (2)$$

Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$F_1\text{-}score = \frac{2 \cdot precision \cdot recall}{(precision + recall)} \quad (3)$$

$F_1\text{-}score$ Is the weighted harmonic mean of precision and recall, also known as the $F_1\text{-}score$.

The symbols in the formula are explained in Table X.

TABLE X
SYMBOL DESCRIPTION

Symbol	Description
Rel Doc	Relevant Documents
Ret Doc	Retrieved Documents

In this research, relevant documents refer to the correctness of the extracted keywords/keyphrases. Retrieved documents refer to keyword candidates extracted by the method.

In order to create a baseline, we hired three master students to evaluate the correctness of the extracted keywords manually. Table IX shows the number of keyword candidates extracted by the traditional methods.

The number of correct keywords is shown in Table X. The number of incorrect keywords is shown in Table XI. The methods include TF-IDF, YAKE, RAKE-NLTK, TextRank4zh, FastTextRank, HarvestText.

The precision, recall, and F-score of the traditional methods are shown in Table XI.

TABLE XI
THE PRECISION, RECALL, AND F1-SCORE OF DIFFERENT METHODS

Methods	Dataset	Precision	Recall	F-score
TF-IDF	A	33.83%	100%	50.56%
	B	52.29%	100%	68.67%
YAKE	A	45.79%	100%	62.81%
	B	60.22%	100%	75.17%
RAKE-NLTK	A	30.98%	100%	47.30%
	B	50.02%	100%	66.68%
TEXTRANK4ZH	A	25.33%	100%	40.43%
	B	42.00%	100%	59.15%
FASTTEXTRANK	A	22.00%	100%	36.07%
	B	28.39%	100%	44.22%
HARVESTTEXT	A	38.77%	100%	55.88%
	B	56.98%	100%	72.60%
CKPE	A	64.43%	100%	78.37%
	B	66.60%	100%	79.95%

It can be seen from Table XI that the effectiveness of the YAKE method for extracting keywords is higher than that of TF-IDF, RAKE-NLTK, TextRank4zh, FastTextRank and HarvestText method. According to the F-score and the Precision values, we can see that the FastTextRank method is less effective than the TextRank4zh method. Because FastTextRank is based on the TextRank4zh method to improve the speed of extracting keywords. Since HarvestText is based on extracting semantic triple to extract keyphrases, its effectiveness is higher than other algorithms, based on statistical rules, location, and other features.

The $F_1\text{-}score$ of YAKE, RANKE-NLTK, TextRank4zh, FastTextRank, HarvestText, and CKPE methods are shown in Fig. 3.

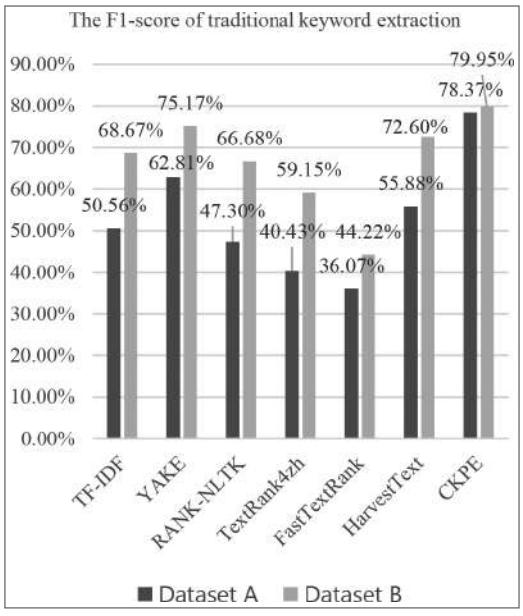

Fig. 3. The F1-score of traditional keyword extraction

TABLE XII
THE PRECISON, RECALL AND F1-SCORE OF THE BERT PRE-TRAINING MODEL

Methods	Dataset	Precision	Recall	F-score
BERT-1	A	45.02%	100%	62.09%
	B	38.73%	100%	55.84%
BERT-2	A	51.98%	100%	68.40%
	B	62.82%	100%	77.16%
BERT-3	A	48.99%	100%	65.76%
	B	62.54%	100%	76.95%
BERT-4	A	34.88%	100%	51.72%
	B	37.32%	100%	54.35%
BERT-5	A	31.84%	100%	48.30%
	B	35.15%	100%	52.02%

It can be seen Table XII the effectiveness of the BERT model for extracting keywords. The effectiveness of the BERT model for extracting keywords varies according to the type of the BERT pre-training model. Among the BERT pre-training model used to extract keywords, the BERT-1 and BERT-2 are more effective. The effectiveness of the BERT-2 model is better than other BERT pre-training models in Table XI. The model of BERT-*N* refers to Table VIII.

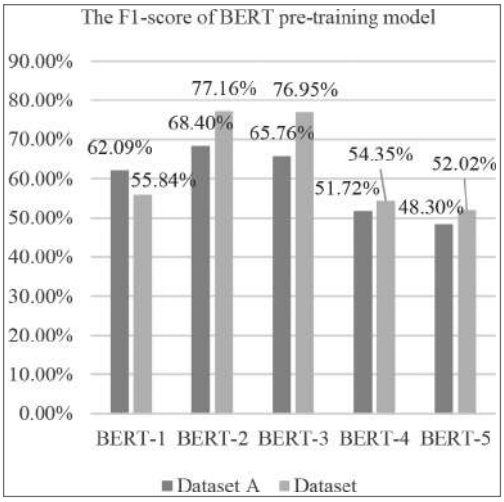

Fig. 4. The F1-score of BERT pre-training model

In addition, we also compared the methods of traditional keyword extraction with the methods of deep learning keyword extraction. The results of the comparison are shown in Fig. 5.

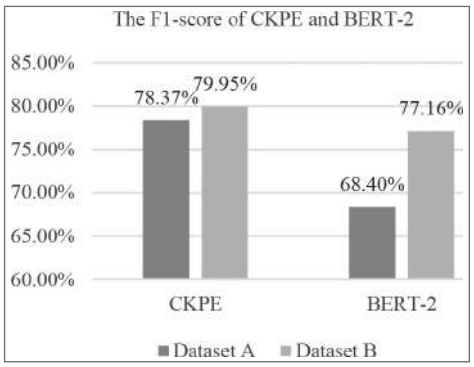

Fig. 5. The F1-score of CKPE and BERT-2

C. *Results of Research*

Among the traditional keyword extraction methods, TF-IDF, YAKE, and CKPE methods have better effects than other methods. Since YAKE extracted long sentences, although the sentence content could help us understand more information on the crowdfunding project, the sentence length was not convenient for later text research. Among the keyword extraction methods based on deep learning, we compared the keyword extraction effect of five BERT embedding models, and the best performance is the BERT-2 model. The results of BERT for Chinese text processing show that the length of the key sentences extracted by the BERT model is too long and some are too short. Sentences that are too long will reduce the computational speed. Sentences that are too short will reduce the validity of the research. Considering the processing speed and validity of the method, we tried to choose a method with a faster calculation speed. Therefore, in this study, we used TF-IDF and CKPE-extracted keywords for external knowledge extraction to obtain features for determining the authenticity of crowdfunding projects [20].

We used machine learning to learn features for classifying fake crowdfunding projects. In the machine learning algorithm, we compared the effectiveness of Decision Tree (DT), Naïve Bayes (NB), Artificial Neural Network (ANN), and Support Vector Machines (SVM) for classifying fake crowdfunding projects. The meta-level and feature selection of the four algorithms and the effectiveness of the optimized algorithms are compared. The specific results are shown in Table XIII.

TABLE XIII
THE RESULTS OF CALSSIFICATION WITH MACHINE LEARNING

Methods	Method	CKPE		TF-IDF	
		A	B	A	B
DT	ML	97.50%	74.22%	97.50%	76.11%
	BE	97.50%	81.56%	97.50%	77.33%
	OP	97.50%	80.52%	97.50%	80.52%
NB	ML	97.50%	74.11%	97.50%	76.22%
	BE	97.50%	79.11%	97.50%	79.44%
	OP	97.50%	83.77%	97.50%	82.11%
ANN	ML	97.50%	77.33%	97.50%	78.22%
	BE	97.50%	80.33%	97.50%	80.33%
	OP	97.50%	81.63%	97.50%	80.61%
SVM	ML	97.50%	80.33%	97.50%	80.33%
	BE	97.50%	80.33%	97.50%	80.33%
	OP	97.50%	80.61%	97.50%	80.61%

Source: ML: Meta-level;
BE: Backward Elimination;
OP: Optimize Parameters (Evolutionary)

Table XIII implies that our proposed method performs better on dataset A, which is because dataset A is a highly imbalanced dataset. Dataset A is from Wikipedia, so we could not achieve its balance between fake and real projects. Therefore, we try to solve this problem by showing with a different dataset which is a more balanced dataset. Dataset B is the balanced dataset we retrieved from Kickstarter and Indiegogo, so the results of our proposed method on dataset B are more indicative of the effectiveness of the method. Among the four methods, DT, NB, ANN, and SVM, we found that the classification algorithm of NB is more effective for the experiment.

Since we determine whether a crowdfunding project is fake or not based on whether the innovative technology proposed by the project can be achieved, the method we propose is only applicable to technology-based crowdfunding projects. In this research, we used the accuracy of human identification as a baseline. Since much fake information about crowdfunding projects is difficult or even impossible to identify artificially, the accuracy of this baseline was 75%.

V. CONCLUSION

There is possible fake information in some crowdfunding projects. Unfortunately, the crowdfunding platform cannot easily identify such fake information in the crowdfunding project. To identify possible fake information in the crowdfunding project, we proposed a method to detect fake crowdfunding projects. In our proposed method, the keyword extraction method is firstly used to extract keywords from the project abstract. Then we search for external knowledge based on keyword extraction, and the feedback information from external knowledge retrieval is used as the feature basis for judging the authenticity of crowdfunding projects. Finally, the machine learning features are used to classify the real and fake crowdfunding projects. We need to extract the keywords of the crowdfunding project and analyze the information of the crowdfunding project. Many studies are devoted to analyzing the successful or failed factors of crowdfunding projects, but they neglect to study the authenticity of crowdfunding projects. In the process of analyzing the authenticity of the crowdfunding project, we need to extract text keywords. At present, there are many methods for extracting keywords from Chinese text. We want to compare the effectiveness of keyword extraction methods for crowdfunding projects through experiments.

The experimental results show that the BERT method is more effective than other methods for keyword extraction. Among the different BERT embedded models, the most effective one is BERT-2. In the classification results, the best result is the NB algorithm, which has an efficiency of 83.77% in the classification of fake crowdfunding projects. In the future, we plan to train the BERT pre-training model to improve the effectiveness of the BERT model in extracting keywords for crowdfunding projects.

ACKNOWLEDGMENT

The first author conducted the experiment and drafted the manuscript. The last author guided and advised the experiment and co-drafted the manuscript. The first and last authors each contributed 50% equally to this work. The last author is the corresponding author.

The first author received scholarship support from CPALL for conducting this research in PIM.

REFERENCES

[1] S. Pandey, S. Goel, S. Bansla et al., "Crowdfunding Fraud Prevention Using Blockchain," in *Proc. 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*. 2019, pp. 1028-1034.

[2] C. Zhang, "Automatic Keyword Extraction from Documents Using Conditional Random Fields," *Journal of Computational Information Systems,* vol. 4, no. 3, pp. 1169-1180, 2008.

[3] R. Campos, V. Mangaravite, A. Pasquali et al., "YAKE! Keyword Extraction from Single Documents Using Multiple Local Features," *Information Sciences,* vol. 509, pp. 257-289, 2020.

[4] C. Wartena, R. Brussee, and W. Slakhorst. "Keyword Extraction Using Word Co-Occurrence," in *Proc. 2010 Workshops on Database and Expert Systems Applications*, 2010, pp. 54-58.

[5] F. Bayatmakou, A. Ahmadi, and A. Mohebi. "Automatic Query-Based Keyword and Keyphrase Extraction," in *Proc. 2017 Artificial Intelligence and Signal Processing Conference (AISP)*, 2017, pp. 325-330.

[6] J. Qu, T. Theeramunkong, N. Le Ming et al. "A Flexible Rule-Based Approach to Learn Medical English-Chinese OOV Term Translations from the Web," *International Journal of Computer Processing of Languages*, vol. 24, no. 2, pp. 207-236, 2012.

[7] R. Mihalcea and P. Tarau, "Textrank: Bringing Order into Text," in *Proc. The 2004 Conference on Empirical Methods in Natural Language Processing*, 2004, pp. 404-411.
[8] S. Rose, D. Engel, N. Cramer et al., "Automatic Keyword Extraction from Individual Documents," *Text Mining: Applications and Theory*, vol. 1, 2010, pp. 1-20.
[9] K. Zhang, H. Xu, J. Tang et al., "Keyword Extraction Using Support Vector Machine," in *Proc. International Conference on Web-Age Information Management*, 2006, pp. 85-96.
[10] J. Qu, NL. Minh, and A. Shimazu, "Web Based English-Chinese OOV Term Translation Using Adaptive Rules and Recursive Feature Selection," In *Proc. The 25th Pacific Asia Conference on Language, Information and Computation*, 2011, pp.1-10.
[11] J. Qu, T. Theeramunkong, C. Nattee et al., "Web Translation of English Medical OOV Terms to Chinese with Data Mining Approach," *Science Technology Asia*, vol. 16, no. 2, pp. 26-40, 2011.
[12] Q. Zhang, Y. Wang, Y. Gong et al., "Keyphrase Extraction Using Deep Recurrent Neural Networks on Twitter," in *Proc. The 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 836-845.
[13] R. Meng, S. Zhao, S. Han et al., "Deep Keyphrase Generation," in *Proc. The 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 582-592.
[14] J. Devlin, M. W. Chang, K. Lee et al., "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in *Proc. The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171-4186.
[15] E. Loper and S. Bird, "Nltk: The Natural Language Toolkit," in *Proc. The ACL Interactive Poster and Demonstration Sessions*, 2004, pp. 214-217.
[16] D. Adimanggala, F. A. Bachtiar, and E. Setiawan, "Evaluasi Topik Tersembunyi Berdasarkan Aspect Extraction Menggunakan Pengembangan Latent Dirichlet Allocation," *Jurnal RESTI*, vol. 5, no. 3, pp. 511-519, 2021.
[17] J. Jiang, "Information Extraction from Text," *Boston, MA: Springer*, 2012, pp. 11-41.
[18] P. Sharma and Y. Li. (2019, Aug. 2). *Self-Supervised Contextual Keword and Keyphrase Retrieval with Self-Labelling*. [Online]. Available: https://shorturl.asia/vwUqZ
[19] J. Qu, N. Phaphoom, C. Wangtragulsang et al., "Social Media Contact Information Extraction," in *Proc. 2018 International Conference on Information Technology (InCIT)*, 2018, pp. 1-5.
[20] Q. Li and J. Qu, "Product Ontology Construction for Crowdfunding Pojects", *presented at the 7th International Conference on Business and Industrial Research (ICBIR) 2022*, Bangkok, Thailand, May 19-20, 2022.

Qi Li is currently studying for the Master of Engineering (Engineering and Technology), Panyapiwat Institute of Management, Thailand. Her research interests are natural language processing and information retrieval.

Jian Qu is a full-time lecturer at the Faculty of Engineering and Technology, Panyapiwat Institute of Management. He received Ph.D. with Outstanding Performance award from Japan Advanced Institute of Science and Technology, Japan, in 2013. He received B.B.A with Summa Cum Laude honors from Institute of International Studies of Ramkhamhaeng University, Thailand, in 2006, and M.S.I.T from Sirindhorn International Institute of Technology, Thammast University, Thailand, in 2010. His research interests are natural language processing, intelligent algorithms, machine learning, machine translation, information retrieval, and image processing.