Current Applied Science and Technology Vol. 23 No. 6 (November-December 2023)

Research article

BM5-SP-SC: A Dual Model Architecture for Contradiction Detection on Crowdfunding Projects

Wenting Hou and Jian Qu*

Faculty of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi, Thailand

Received: 26 April 2022, Revised: 25 August 2022, Accepted: 24 March 2023

DOI: 10.55003/cast.2023.06.23.007

Abstract

Keywords

crowdfunding projects;

contradiction detection;

knowledge extraction;

BERT;

MT5;

feature information extraction

Despite the prevalence of scams in crowdfunding projects, currently, there is little research into the identification of fraudulent or infeasible crowdfunding projects. Since detecting fraudulent crowdfunding projects is challenging, most existing research on fake information has focused on detecting fake news or fake charity crowdfunding projects based on social media, but research on fraudulent or infeasible crowdfunding projects is very limited. Therefore, to solve this problem, we focus on how to detect fraudulent crowdfunding projects based on knowledge extraction and contradiction detection. We proposed a novel method called BM5-SP-SC (BERT-MT5-Sentence Pattern-Sentiment Classification). BM5 (BERT-MT5), which is built from a combination of a key-BERT and a fine-tuned MT5 transformers, was used to extract feature information from crowdfunding projects. We proposed a novel method for MT5 training to construct an adaptive BM5 model. The correct rate of keywords extracted by our novel adaptive BM5 model was up to 72.7%, the recall was 100%, and the F-measure was up to 84.19%. The minimum train loss of the BM5 model was up to 0.1342, and the evaluation loss achieved was 0.3064. The BLEU score of summary-to-keyword was 37.336. Moreover, we proposed an SP (Sentence Pattern) matching method to achieve knowledge extraction. Furthermore, SC (Sentiment Classification) was used to build a sentiment classifier thesaurus for identifying fraudulent and infeasible crowdfunding projects. Our proposed BM5-SP-SC achieved an overall accuracy of 85.26% in detecting fraudulent crowdfunding projects.

*Corresponding author: Tel.: (+66) 863759307
E-mail: jianqu@pim.ac.th

1. Introduction

Crowdfunding refers to a mode in which companies or individuals showcase their creativity to the public through the Internet to gain attention and draw investment. We found that the failure rate of crowdfunding projects in 2015 was 59.55% [1], while in 2019, it reached 64.62% [2], which means that only 35.38% of the projects were successful in crowdfunding. We can clearly see that the failure rate of crowdfunding projects is already very high, and it keeps rising.

In order to solve this problem, we must first understand why these projects failed. Failed projects are further divided into feasible projects that ended up failing, and fraudulent crowdfunding projects that were not supposed to be successful to begin with. The latter group of fraudulent crowdfunding projects can be divided further into fraudulent yet feasible projects (i.e., could be accomplished if done in good faith) and fraudulent and impractical projects (Figure 1). The first case is a fraudulent feasible project. In this group, there is no problem with the project and if done in good faith the project could have been done. However, the project sponsor may not faithfully invest the funds in the project after the funds have been raised, but instead embezzles the funds for private use, and then declares the project has failed for other reasons. The second case is a fraudulent and impractical project. The project is impossible to complete, but fundraisers deceive investors by making videos and pictures, making investors mistakenly believe that the project is feasible. For example, Indiegogo launched a project called Triton, and it claimed that people could stay in the water for about 45 min on a very small and lightweight respirator. Currently, divers still have to carry heavy tanks and assorted equipment and have to consider all kinds of inconveniences and hazards caused by gas cylinders (like running out of air, or like getting stuck while diving in a cave). So, this project was like a dream for many divers. The creativity involved in the project may be excellent, and it may have looked very promising. However, the Triton faced limitations in battery design, high-pressure storage, and filtration that were not likely to have been overcome for decades. Dr. Alistair Dove, the former associate editor of Deep Sea News, estimated that it would require 90 L of water per min to generate enough air for human. This was similar to a 1/4 horsepower sump pump, and this was all supposed to be done by a small lithium-ion battery. Considering power consumption, this setup was not likely to work, and it could have been seen as an imitation of a similar gadget used in Star Wars and 007 movies. Secondly, drawing just one breath would have taken about 2 L of air while exercising (or about 0.5 L at rest). Therefore, the system would have needed a reservoir to hold the compressed air. They claimed that a tiny built-in reservoir would hold it, which was impossible. Breathing like a fish sounds logical, but it is technically impossible. However, it was also challenging for ordinary people to prove that this project was not feasible or 'debunk' it. Many diving enthusiasts did not find any problems and supported the project.

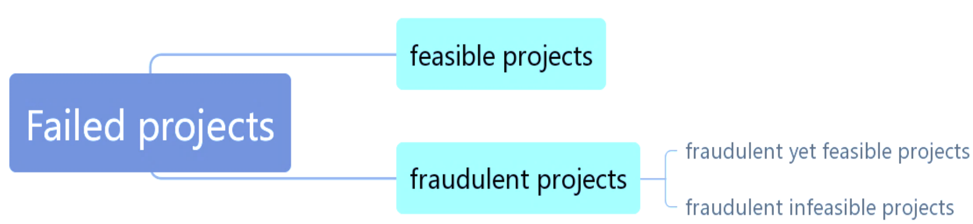

Figure 1. Classification of failed projects

Crowdfunding platforms have few, if any, automated ways to detect fraudulent, infeasible projects. This research is a first step towards investigation of inconsistency/contradiction detection for crowdfunding projects. Existing research work on fake information has been focused on fake news on social media. Wang [3] proposed a new, publicly available data set called LIAR for fake news detection. However, LIAR's facts and short real-world statements are judgments of the correctness of the news content, not logical feasibility exploration. So, if this approach were applied to our task as it is, there would be no statements about newly published crowdfunding projects in LIAR. Therefore, logical contradiction could not be detected. De Marneffe *et al*. [4] proposed a contradiction definition of NLP tasks and developed a usable corpus to construct contradictory typology. They divided contradictions into two main types: (1) Contradictions in antonyms, negations, and date/number mismatches, which were easier to find. However, these contradictions were based on common sense, and our tasks cannot always be solved with common sense. (2) Contradiction arising from the use of facts or modal verbs, structure and subtle vocabulary comparisons, and World Knowledge (WK). They claimed that it was very difficult to detect the type (2) contradiction with their system. There is limited research on fraudulent crowdfunding projects on Google Scholar. Only one paper about fraud detection in crowdfunding projects could be found. In this paper, Perez *et al*. [5] used a social media method to analyze data collected from different crowdfunding platforms and flagged 700 campaigns as fraudulent. They used social media and social credibility to determine whether projects were fraudulent, and the precision rate was as high as 90.14%. However, this paper only focused on donations to charity-based crowdfunding environments. They used social network-graph sentiment retrieval. We applied this approach to our task and extracted keyword features. However, our task involves many aspects such as technology, games, and fiction, which can result in a low recall of the results. Therefore, we decided to explore relatively uncharted territory to detect logical contradictions in fraudulent infeasible projects. There are three steps in our research: information retrieval, knowledge extraction, and contradiction detection. We proposed a novel method for training MT5 in the first step and a novel sentence pattern matching method in the second step to achieve knowledge retrieval. Finally, we built a sentiment classifier thesaurus to detect contradictions through sentiment classification according to extracted knowledge.

There are many methods for extracting keywords [6]: TFIDF [7, 8], TextRank [9, 10], NLTK [11, 12], LIAAD [13-15], Harvest [16, 17], BERT [18, 19] and MT5 [20]. Sun *et al*. [19] proposed a text keyword extraction method based on BERT and compared it with TFIDF, TextRank, and LDA. Based on 300 scientific papers downloaded from the Wanfang database, TFIDF achieved an F-measure of 36.4%, TextRank achieved an F-measure of 40.7%, LDA achieved an F-measure of 42.0%, and a keyword extraction algorithm based on BERT, and multi class feature fusion achieved an F-measure of 43.6%. The results show that the combination algorithm based on BERT was better than the single extraction algorithm. In our research, we used the word segmentation of BERT to propose a new method for training MT5. We called this method BM5. This new method can achieve keyword extraction and selection of the crowdfunding project at the same time.

Qu *et al*. [21] proposed a new integrated sorting method for definition retrieval. Statistically speaking, most of the retrieved snippets do not contain definitions. However, snippets with definitions do have some patterns, e.g., 'Mae West was an American actress.' According to observation, some verbs can be used to find snippets with definitions. Qu *et al*. [22] recommended using rule-based sorting methods to define the rank of snippets. If applied to our task, the definition of keywords cannot get contradictory information, so we proposed a sentence pattern matching method to get contradictory information to detect fraudulent projects through sentence modification retrieval of extracted keywords. We proposed five sentence structures and selected the best performing sentence structure for knowledge retrieval.

As for contradiction detection, we divided contradiction detection reviews into fake news detection and non-news contradiction detection. Automatic fake news detection is a challenging problem in contradiction detection, with a huge real-world political and social impact. However, due to the lack of labeled benchmark data sets, the ability of statistical methods to combat fake news have been greatly restricted. Wang [3] proposed a new, publicly available data set for fake news detection. LIAR is a new data set for automatic fake news detection. LIAR's facts and short real-world statements come from different backgrounds and speakers, making it possible to develop a wide range of fake news detectors. The researcher aimed to point out promising research directions from the perspective of data mining and outline four directions of research: data-oriented, feature-oriented, model-oriented, and application-oriented. For data-oriented, four public fake news detection data sets were compared: BuzzFeedNews, LIAR, BS Detector, and CREDBANK, and the features extracted from each data set were highlighted. It was proved that no existing public data set could provide all possible relevant features, and proposed to create a comprehensive and large-scale fake news benchmark data set that could be used by researchers to promote further research in this field. However, if this approach were applied to our task in its usual form, there would be a drawback: there are no statements about newly published crowdfunding projects in the data set. For feature-oriented, the researcher stated that there were two primary sources of data: news content and social background. This research direction was not focused on logical contradiction; its focus was on the background of the project sponsor, which was not the focus of our research. For model-oriented, examples include Naive Bayes, Decision Trees, Logistic Regression, and support vector machines (SVM), most of which are supervised. Shu *et al*. [23] also proposed that limited or unlabeled fake news fragments could be used in scenarios where semi-supervised or unsupervised models were applied. For application-oriented, application-oriented fake news research covers research in other fields besides fake news detection. Furthermore, Shu *et al*. [23] also proposed two main directions: fake news proliferation and fake news intervention. However, there is a big difference between fake news and fraudulent projects, frequency can be a crucial indicator for fake news detection, but such a method cannot be applied for fraudulent projects.

Non-news research related to contradiction detection included the work of De Marneffe *et al*. [4] on finding contradictions in texts. They proposed the contradiction definition of NLP tasks and developed a usable corpus to construct contradictory typology. For example, if the words in the two paragraphs are antonyms, the sentences are similar, especially in terms of polarity, and conflicts will occur. However, the keywords in our task are generally free of such contradictions. If this approach were applied to our task in its original form, the recall of contradiction would drop.

To address above problems, we proposed a novel adaptive BM5-SP-SC. Contradiction detection on crowdfunding projects can be achieved with following steps: snippet retrieval, keyword extraction, knowledge extraction, and contradiction detection. Firstly, we used Google API to retrieve information automatically. Secondly, keyword was extracted from such information using our novel adaptive BM5 model, which can generate and select possible correct keywords using an adaptive fuzzy training method for Google MT5. Thirdly, we used the extracted keywords as input to combine a group of novel regular expression rules to find possible correct knowledge from World Wide Web using our novel SP (Sentence Pattern). Fourthly, we conducted emotional analysis on the acquired knowledge using our modified SC (Sentiment Classification) to detect contradiction.

2. Materials and Methods

This research was divided into four major steps: snippet retrieval, keyword extraction and selection using BM5, knowledge extraction, and contradiction detection, as shown in Figure 2.

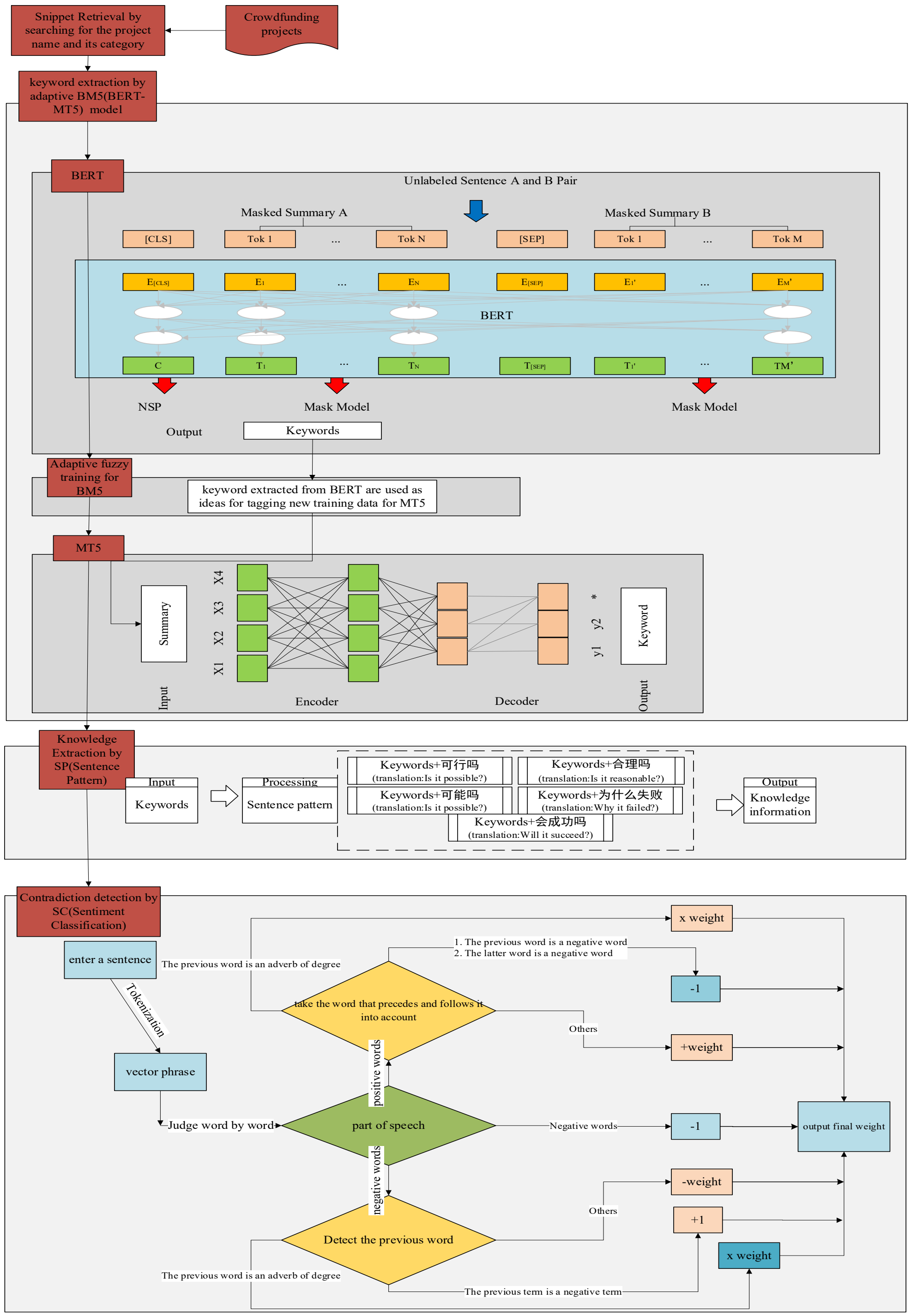

Figure 2. Overall flow chart of the proposed method

2.1 Snippet retrieval

As the name of a crowdfunding project may be ambiguous, the information that we search for may not be relevant to that project. For example, if you search for the "EOS" project, there may be a person or a song called "EOS". Therefore, we had to find an effective way to reduce the ambiguity of the search text. There are four ways to construct the search text: search for the project name, search for the project name and its category, search for the project name and its platform, and search for the project name plus category plus platform [24]. This is shown in Table 1. We refer to A as the "project name", B as the "project category" and C as the "crowdfunding platform". We conducted an experiment to find out which search method was better. We searched Wikipedia for the top 20 projects in the list of the highest funded crowdfunding projects. Ten snippets were intercepted for each project, and 200 snippets were intercepted for 20 projects. The retrieved summaries were different for each search method; therefore, we got a total of 800 different summaries. We used the 800 summaries to compare the validity rates of the four search methods. We judged the correctness of each summary. When the input was "A", the number of correct snippets was 143 and the effective ratio was 71.5%. By examining 800 summaries, we found that searching for the project name and its category to retrieve the summaries was the most effective and least intrusive way to construct the query text. The validity ratio is ranked from highest to lowest, as shown in Table 2. In summary, in this research, we chose to use "A+B" as the input for the snippet retrieval.

Table 1. Details of four different search methods

Search Method	Instance	Shortening
"name+category"	"Star Citizen+电子游戏"	"A+B"
"name"	"Star Citizen"	"A"
"name+category+platform"	"Star Citizen+电子游戏+Kickstarter"	"A+B+C"
"name+platform"	"Star Citizen+Kickstarter"	"A+C"

Note: A = name of the project itself, B = category of the project, C = crowdfunding platform of the project

Table 2. Effective ratio of different search methods

Input	Number of Valid Snippets	Effective Ratio
"A+B"	180	90%
"A"	143	71.5%
"A+B+C"	129	64.5%
"A+C"	115	57.5%

There were 100 crowdfunding projects, and we separated the obtained web text into three different fields: URL, title, and summary. An example of the snippet retrieved is shown in Figure 3, and we performed keyword extraction on the summary.

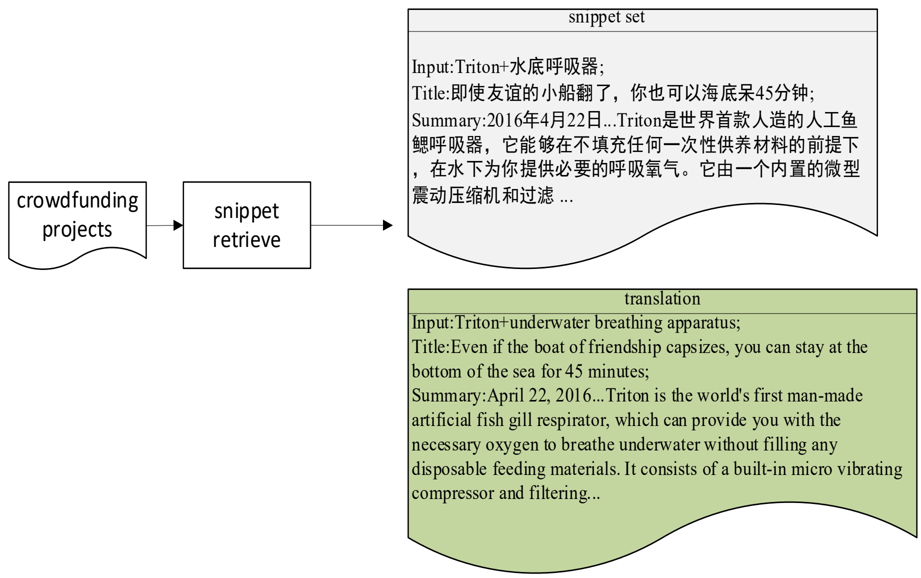

Figure 3. Flow chart of the snippet retrieval

2.2 BM5 model

We proposed a novel method with a construction connecting two transformers, which were BERT for keyword extraction and MT5, where BERT for keyword extraction is a sequence classification model, and MT5 is a text-to-text transformer model. In our previous studies [24], we had compared seven keyword extraction methods and found that one of the BERT keyword models performed better than other keyword extraction methods on crowdfunding project datasets. We used keyword extraction to understand the main elements of the project, which were used to further determine the feasibility of the project. Therefore, we used the keywords extracted by BERT as the data for fine-tuning the MT5 model, as shown in Figure 2. We also proposed a novel adaptive fuzzy training method for BM5 (BERT + MT5) model, and this novel adaptive fuzzy training method was explained in Figure 4.

The reason for choosing BERT is tested. Two databases were used for this research: List A, which was a list of the highest funded crowdfunding projects on Wikipedia (only information from the Kickstarter and IndieGoGo platforms was retained for better comparison with List B), and List B, which was used for testing the performance of our proposed method. List A was publicly available data, which was only used to derive which keyword extraction method was more effective in terms of crowdfunding projects. We performed a keyword extraction method on the 5340 summaries retrieved from the 120 projects [24] in List A. For these 5340 summaries, we generated 42,825 keywords using NLTK, LIAAD and Harvest, and another set of 64,044 keywords using the four BERT-based models. We hired five graduate students to help us mark these keywords for correctness. Finally, we created a corpus to store those 106,869 pairs of texts and their keywords that were tagged for keyword extraction based on crowdfunding projects. There were 120 projects in total. For each project, 100 summaries were retrieved. We took 5,340 summaries as input and set the number of keywords extracted by each method to 3. Thus, NLTK generated 15,915 keywords, LIAAD generated 15,626 keywords, Harvest generated 11,284 keywords and BERT generated 16,011 keywords. As shown in Table 3, the results show that the BERT model worked better because the selected candidate words were closer to the meaning of the document. We saw that NLTK has the lowest F-measure at 15.6%. BERT has the highest F-measure at 28.6%. It

```
Input: Summaries of snippets retrieved from the Internet, Ss; Names of crowdfunding
projects, Cpn.

Output: Keyword extracted by MT5, Km; Keyword extracted from summaries of
snippets retrieved from the Internet, Kss; Fixed sentence "This summary is invalid", Fs.

For each Ss do
  If (Cpn found in Ss){
    Km = Kss;}
  else{
    Km = Fs;}
End
```

Figure 4. Adaptive fuzzy training for BM5

Table 3. Comparison of different methods of keyword extraction

Method	Total number of keywords	Effective number	Precision	Recall	F-measure
NLTK	15915	1349	8.5%	1	15.6%
LIAAD	15626	1984	12.7%	1	22.5%
Harvest	11284	1533	13.6%	1	23.9%
Bert(V1)	16011	2676	16.7%	1	28.6%

outperformed the other three methods, and importantly BERT was the only one of the four keyword extraction methods that could obtain a bidirectional feature representation of the context. For example, the word "address" has different meanings in different contexts: "The recommendation letter has been sent to your address" and "We need a leader to address COVID-19 on a global scale". You need to look at the sentence as a whole to understand its syntax and semantics. The BERT model embedding looks at the whole sentence to understand its syntax, semantics, and context, thus improving the precision of the NLP task. Therefore, BERT is the most suitable method for extracting keywords for crowdfunding projects.

According to our observations, the summaries of snippets retrieved from the Internet for many crowdfunding projects had their own characteristics. Statistically speaking, most of the retrieved summaries contains non-useful information, and were thus referred to as invalid summaries. The problem was how to distinguish between valid and invalid summaries. However, valid (snippets contains useful information) summaries have some patterns. From our observations, the names of the crowdfunding projects could be used to find possible valid summaries.

The development of adaptive fuzzy training for BM5 is shown in Figure 4. Let *Ss* be the summaries of snippets retrieved from the Internet, *Cpn* be the names of crowdfunding projects, *Km* be the keyword extracted by MT5, *Kss* be the keyword extracted from summaries of snippets retrieved from the Internet, and *Fs* be the fixed sentence "This summary is invalid". For each *Ss* term: if the *Cpn* is found in the *Ss*, we consider the summary as valid and take the keyword extracted from MT5; however, if *Ss* does not contain *Cpn*, we consider the summary as invalid and take the fixed sentence "This summary is invalid" as the keyword.

The fuzzy training method also reduced the timing for data tagging. The detailed flowchart of our method is shown in Figure 5.

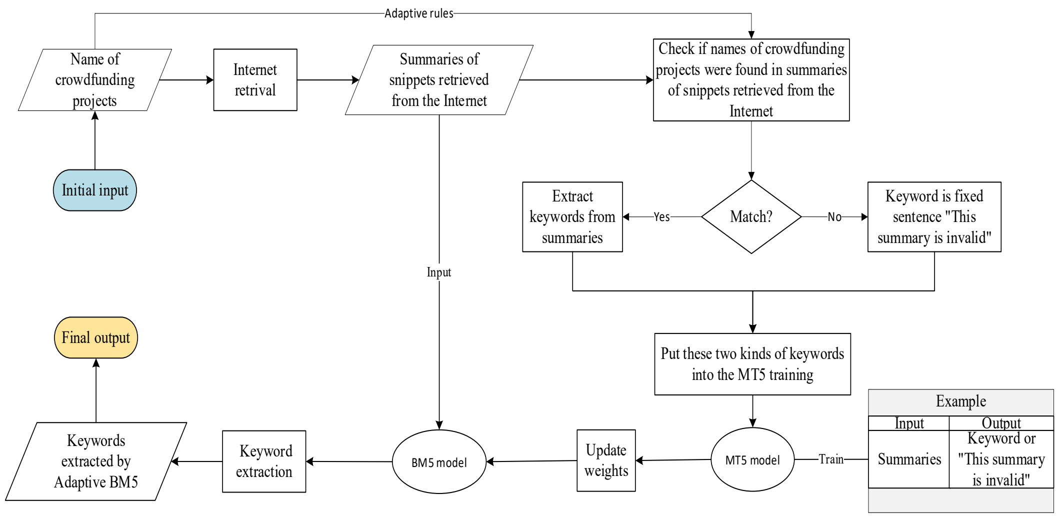

Figure 5. The detailed flowchart of adaptive BM5

In our research, the input of BERT was summary (snippet), and the output of BERT was keywords. We found that BERT was not perfect for keyword segmentation. To fine-tune MT5, keywords extracted from BERT were used as ideas for tagging new training data of MT5. The input of the model was the summary, and the output of the model was the adjusted keywords, as shown in Figure 6. The keywords were then assessed for suitability, as shown in Figure 5. If the keywords provided invalid information according to our adaptive fuzzy training method, such keywords were removed and replaced with ‘此 summary 无效’ (translation: this summary is invalid). For the keyword providing valid information, we would keep the keyword as-is. This novel method for training MT5 was called adaptive BM5. Our proposed novel adaptive BM5 model was able to accomplish keyword extraction and selection simultaneously.

2.3 SP (Sentence Pattern)

Other researchers have often dealt with definition retrieval, but definition retrieval is more about interpreting nominal definitions [21]. For our research, keywords are more about the function of expressing projects, which is done in a sentence, and not in a phrase. We proposed a novel query sentence construction method to extract knowledge, as shown in Figure 7. The five sentence patterns shown are the most common synonyms used in Chinese, and these five sentence patterns are the most probable ways of asking whether the project will be successful or not. When we used the proposed SP (Sentence Pattern) for knowledge retrieval, three kinds of knowledge information could be retrieved: 1) A detailed description of the project, which describes the positive introduction of the project on the official website, without any contradictory information. 2) The reasons why the project failed, such as financial problems, false propaganda, and failure to deliver on time; these problems are not the contradictory information about the project logic that we want. 3) The reasons for the contradiction to explain that the project is impossible, which is what we need. The only downside to this method is that if the project has not been discussed online, it is difficult to judge whether it is a fraudulent project using this method. However, our goal is to gather different opinions from users in various fields, so even if this approach is flawed, it is possible to obtain supporting information.

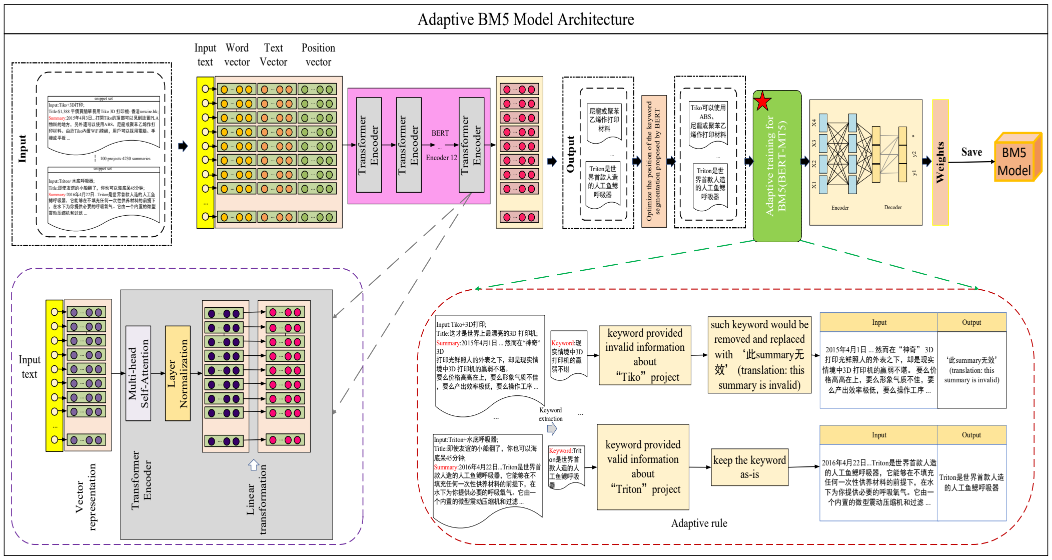

Figure 6. The model architecture of adaptive BM5

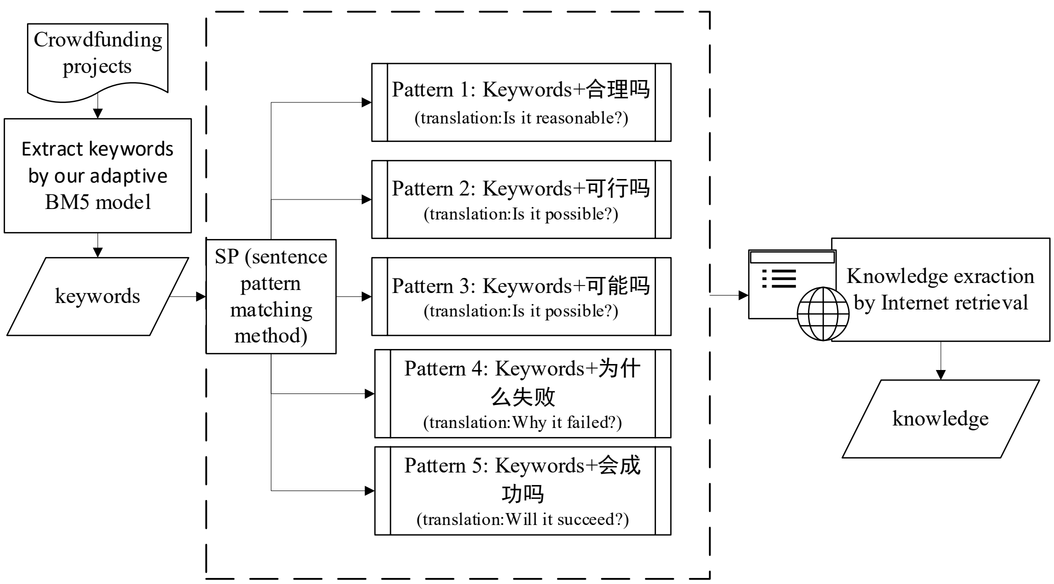

Figure 7. Flow chart of the knowledge extraction

2.4 SC (Sentiment Classification)

To achieve contradiction detection, we explored the information from the retrieved websites. We found that a website could expose project contradictions, words like 'failure', 'impossible', and 'self-contradiction' would frequently appear in the text. We constructed a list of emotional words containing positive and negative words for crowdfunding projects and used it to achieve emotional classification, as shown in Figure 8. We adopted a dictionary-based sentiment analysis method, and the emotion polarity of the text was judged by the emotion score. As shown in Figure 9, if the output final weight was greater than 0, it was judged as positive thus indicting a feasible project; if the output final weight was less than 0, it was judged as negative thus indicating a fraudulent infeasible project.

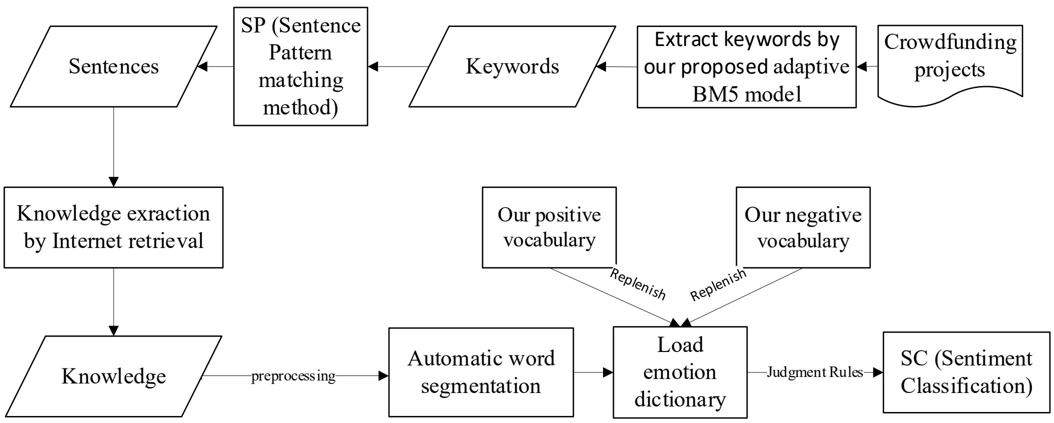

Figure 8. Text sentiment classification based on sentiment dictionary

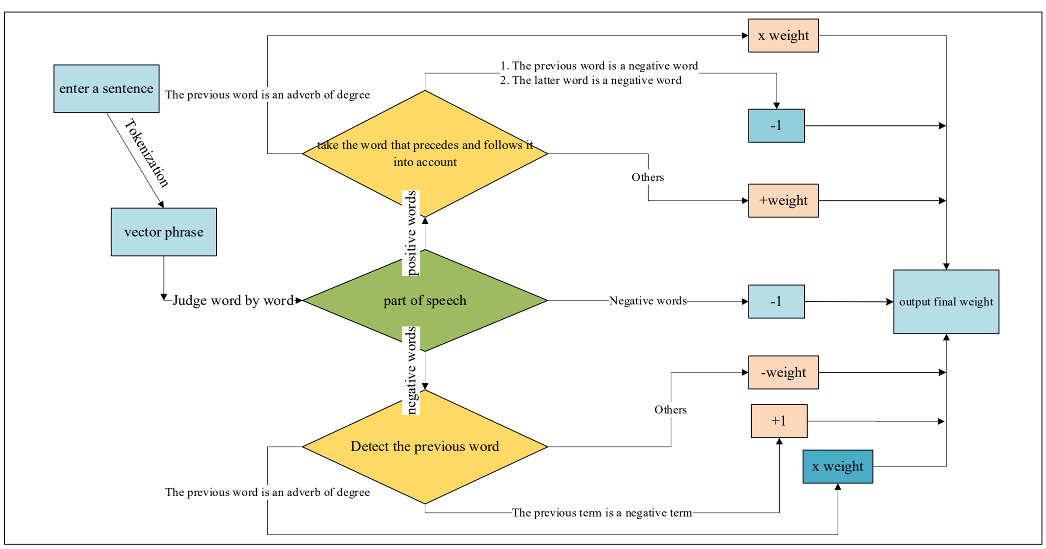

Figure 9. Flow chart of our sentiment classification

3. Results and Discussion

3.1 Data collection

The original data was a list of the highest-funded crowdfunding projects on Wikipedia named List A. List A was publicly available data, which was only used to derive which keyword extraction method was more effective on crowdfunding projects, but there were so few fraudulent infeasible projects in list A, which created a serious data imbalance problem, so we reconstructed a relatively balanced dataset named List B. We manually collected 100 projects according to famous ranking of the crowdfunding platform (Kickstarter and IndieGoGo platforms), of which 20 projects were fraudulent and infeasible. The remaining 80 projects were feasible projects. We searched 100 snippets for each project, which were stored in the database, as shown in Table 4. Since some projects had very little information on the Internet, so they potentially offered less than 100 snippets. In the end, there were 4,230 snippets, including 317,707 characters. Figure 10 shows data sizes for each step.

Table 4. An example of database for snippet retrieval

ID	Keyword	URL	Title	Summary
1	Triton+水底呼吸器	http://www.sohu.com/a/285416143_120078003	猫眼：“鱼鳃”水下呼吸器_潜水	2014年8月19日 ... 据设计师介绍，为了解决这个问题，Triton 不会…

3.2 Evaluation matric

In this section, we used precision, recall, accuracy, F-measure and BLEU to evaluate our method. The formulas of precision, recall, accuracy, F-measure, and BLEU are given in equations (1), (2), (3), (4) and (6). The symbols in the formula are explained in Table 5.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

Precision is the ratio of predicted correct and real feasible projects to predicted feasible projects.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

Recall is the proportion of predicted correct and real feasible projects to real feasible projects.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

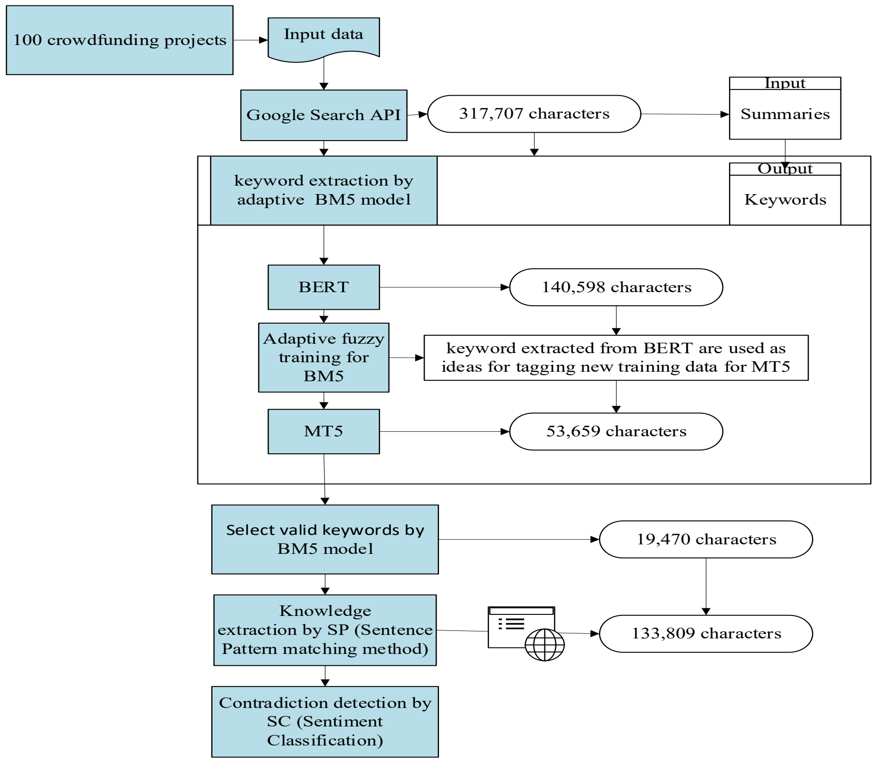

Figure 10. Data size for each task

Table 5. Symbol description

Symbol	Description
TP	predicted correct and feasible project
FN	predicted incorrect and infeasible project
TN	predicted correct and infeasible project
FP	predicted incorrect and feasible project

F-measure is the weighted harmonic mean of precision and recall.

$$BP = \begin{cases} 1, \text{if } c > r \\ e^{(1-r/c)}, \text{if } c \leq r \end{cases} \tag{5}$$

Where c is the length of the candidate keyword and r is the length of effective reference corpus. We chose the brevity penalty to be a decaying exponential in r/c, meaning $e^{(1-r/c)}$. We computed the brevity penalty BP and then calculated the BLEU score. The formula for calculating BP is shown in equation (5).

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} W_n \log P_n\right) \tag{6}$$

We first computed the n-gram that matched sentence by sentence. Next, we added the clipped n-gram counts for all the candidate sentences and divided by the number of candidate n-grams in the test corpus to compute a modified precision score, *Pn.*

N is a constant, in our baseline, and we used N = 4 and uniform weights *Wn* = 1/N [25]. BLEU (Bilingual Evaluation Alternate) is a score that compares a candidate translation of a text with one or more reference translations. It is widely used for evaluating NLP tasks. In this research, BLEU score was used for evaluating and scoring candidate keywords generated from the BM5 model.

3.3 Methods for evaluating BM5

The quality of a model can be evaluated in two parts. Firstly, the loss value from model training. The training loss value can reflect the quality of the model architecture to a certain extent. Secondly, the BLEU score and precision of the actual keyword extraction. Although the loss value was generated in the model training part, we could not judge the quality of the model based on the training loss value alone. So, after model training, we saved the best model and used it to test the evaluation set. Moreover, to avoid data overfitting, we also employed 10-fold cross validation to better evaluate our proposed BM5 model.

3.4 Experiments for BM5

We used 10-fold cross validation to evaluate the model performance. We randomly divided the data set into 10 parts, took turns to train 9 of them and verified 1 of them, and took the mean value of the results of 10 folds as an estimate of the precision of the algorithm. The precision of keywords was 72.7%. However, without using 10-fold cross validation, a precision of 94.89% could be achieved if 90% were used for training data, and 10% were used for testing data, as shown in Figure 11.

There were 4,230 summaries, and 4,014 keywords extracted by the BM5 were valid. There are 317,707 characters as input, and 53,659 characters as output. The precision of keywords was up to 72.7%. Our proposed method is a suitable method for extracting keywords from crowdfunding projects. The most important reason is that MT5 reframes any NLP task as a text-to-text task, which means that both the input and the output are text sequences. The minimum train loss of our proposed model can reach 0.1342, as shown in Figure 12. Evaluation loss can achieve 0.3064, as shown in Figure 13.

The standard metric used to evaluate and compare machine translation models is the BLEU score, we also used it to evaluate the sequence-to-sequence task performance in this research. Our input are the summaries of projects retrieved from the Internet. Output is the keywords extracted from BM5. The BLEU [25] score of summary-to-keyword is 37.336.

3.5 Experiment on Sentence Patterns

In order to test which of these five sentence patterns could really extract contradictory information, we used the keywords from 20 randomly selected fraudulent infeasible projects and 10 feasible projects to test the applicability of these five sentence patterns for knowledge extraction. Since this step was mainly to test the applicability of these five sentence patterns to detect the fraudulent infeasible projects, we controlled the ratio to 2:1 (more fraudulent infeasible projects than feasible projects). For these 30 projects, the web returned 1180 summaries, BM5 generated 3549 keywords; however, only 90 keywords were valid, and therefore a maximum of 900 summaries could be retrieved using knowledge extraction sentence patterns. Five sentence patterns each retrieved less than 900 summaries, which is shown in Table 6.

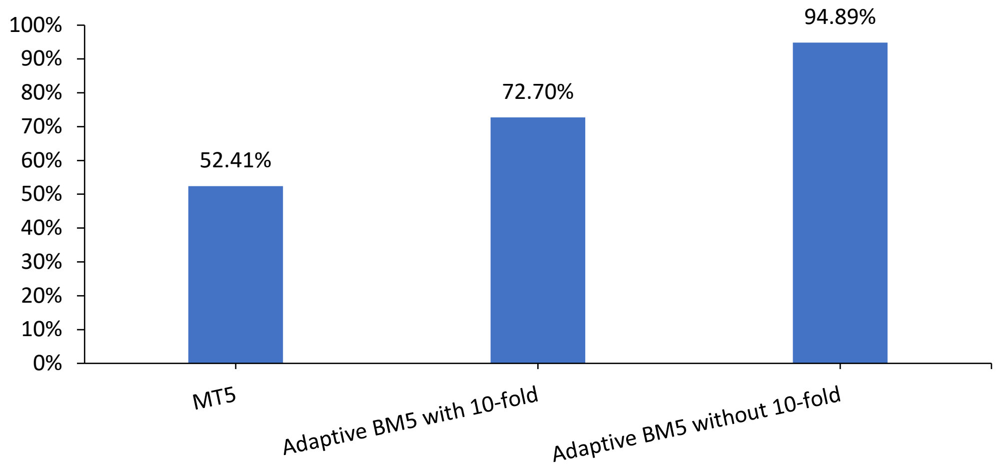

Figure 11. Comparison of training methods

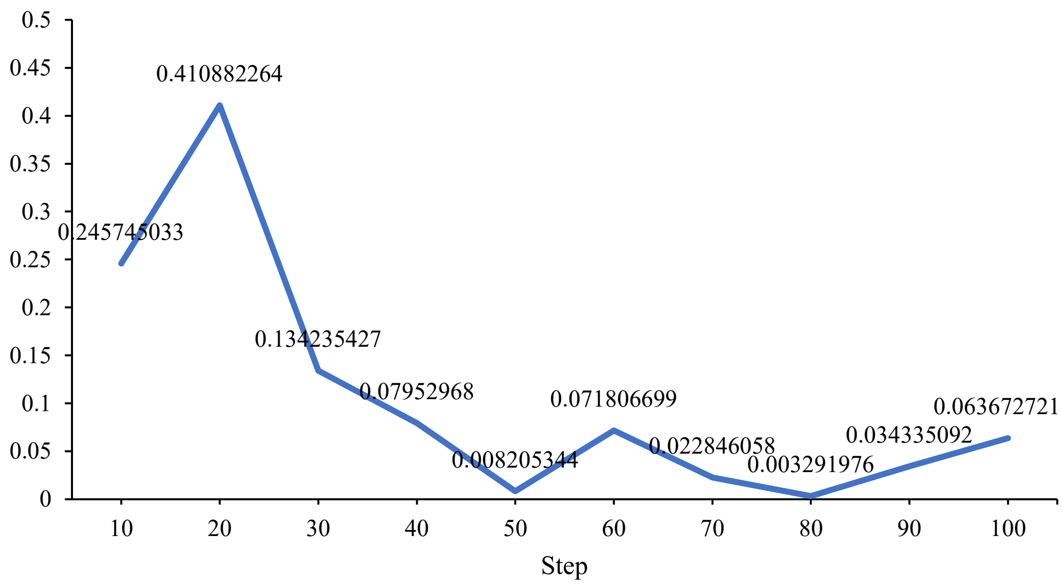

Figure 12. Result of train loss

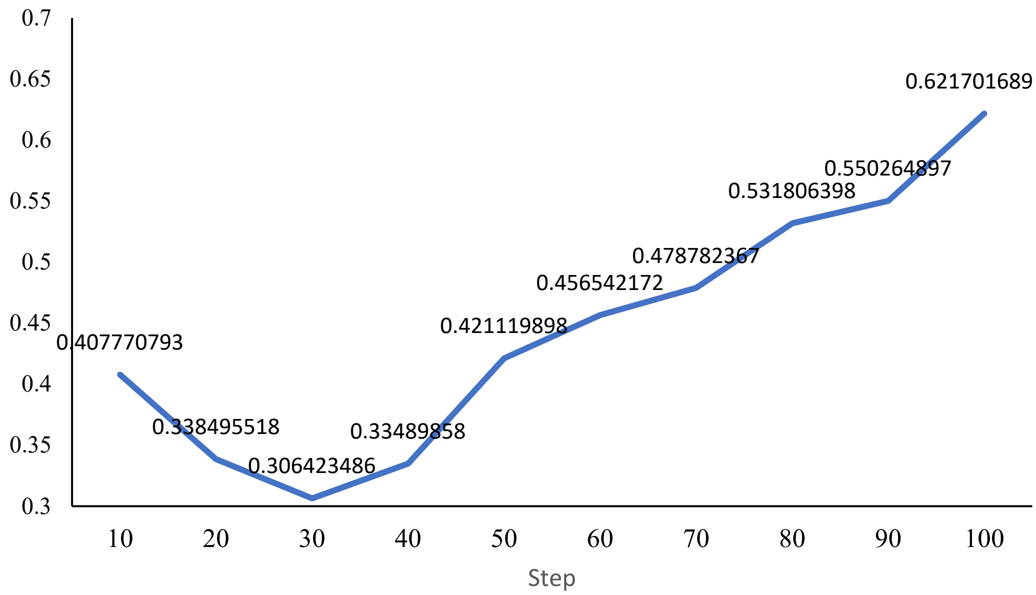

Figure 13. Result of evaluation loss

Table 6. The recall rate of different sentence patterns

Pattern	Number of Summaries	Recall
Pattern 1	821	91.22%
Pattern 2	820	91.11%
Pattern 3	830	92.22%
Pattern 4	697	77.44%
Pattern 5	784	87.11%

3.6 Experiments on possible valid websites for contradiction information

According to existing studies, we tested the detection of contradictory information using URL classification and website type [21], but in the end, only sentiment detection of web page content was able to detect possible contradictory information. Moreover, this method returned a lot of useless information, thus increasing noise in the data. In the knowledge extraction step, we found that the higher ranked websites were more relevant to the project, so we tested how many websites could be selected to detect contradictions with the least noise using 20 fraudulent infeasible projects. The 20 fraudulent infeasible projects produced 601 summaries on the web, generated 1803 keywords, with only 125 valid keywords in the end. When query of sentence pattern was performed to extract knowledge, ideally 1250 websites could have been generated if the first 10 websites were selected, but due to insufficient feedback from some keywords, only 1153 websites were generated, resulting in only 25 valid URLs. We found that there was too much noise, so we tested various valid websites, and when TOP5, TOP4, TOP3, TOP2 and TOP1 websites were selected, we found that the most accurate and least noisy was the TOP2 websites. We therefore chose TOP2 websites to extract contradictory information, as shown in Table 7.

Table 7. Comparison of top N websites in accuracy

Top N	All websites	Valid websites	Accuracy
Top 10	1153	25	2.2%
Top 5	600	18	3%
Top 4	484	15	3.1%
Top 3	367	14	3.8%
Top 2	247	12	4.9%
Top 1	126	6	4.8%

3.7 Experiments for fraudulent project detection

Out of the 100 projects, only 95 projects were detected in the end because 5 projects had no valid keywords on the Internet (data as of 1 December 2021). The accuracy rates of different methods are shown in Table 8. For a better comparison with human results, we used accuracy uniformly to compare between different methods for detecting fraudulent projects. The accuracy of human result was up to 91.25%, but machine learning is not as good as human when it comes to text understanding, the accuracy of our proposed method for detecting fraudulent crowdfunding projects was 85.26%. The existing method proposed by Perez *et al*. [5] only achieved an accuracy of 21.32%.

Table 8. Comparison of detecting fraudulent projects by different methods

Method	Accuracy
BM5-SP-SC	85.26%
Human	91.25%
Existing social network method (Perez B)	21.32%

Perez *et al*. [5] used a social network method to analyze charity crowdfunding projects and flagged 700 projects as fraudulent. They used a credibility score of social media and social networks to determine if such projects were fraudulent, and this social network method produced an accuracy of 90.14%. However, our task involved technology and gaming crowdfunding projects, and the application of the social network method would have resulted in low recall.

For fraudulent infeasible crowdfunding project detection, our method achieved a precision of 92.21%, a recall of 89.87% and a F-measure of 91.02%. Details of project prediction are shown in Table 9.

Table 9. Confusion matrix of project prediction

	Feasible Project (P)	Infeasible Project (N)
Predicted correct (T)	71	10
Predicted incorrect (F)	6	8

4. Conclusions

This research proposed a novel method for detecting fraudulent crowdfunding projects named BM5-SP-SC. For extracting feature information, we proposed a novel method with a combination of BERT and MT5, called BM5, and compared it with NLTK, LIAAD, Harvest, and BERT. Our proposed method performed better than existing methods for extracting feature information from crowdfunding projects with the F-measure of 84.19%. The minimum evaluation loss of BM5 was

0.3064. We proposed a novel adaptive fuzzy method for training BM5. We also proposed a novel sentence pattern matching method named SP to extract knowledge information, which achieved a recall of 92.22% for the knowledge information retrieval. We also built a sentiment word list for crowdfunding projects and used it to implement sentiment classification for detecting fraudulent crowdfunding projects. We called this approach SC. Our proposed method, called BM5-SP-SC, achieved an overall accuracy of 85.26%, while the existing method could only achieve an accuracy of 21.32% for fraudulent crowdfunding project detection.

In our research, there are still some projects for which no valid information can be found from the web retrieved summaries. In fact, fewer effective summaries result in fewer effective keywords, but as long as each project can produce some keywords for knowledge extraction, we may get some information from the project. The only disadvantage, however, is that if the project has no information on the Internet besides the crowdfunding platform and no one is discussing it, our method is less effective. In the future, we hope to analyze information on the official web-pages of the crowdfunding project together with web retrieved summaries, which might enable us to detect more contradictory projects by fusing more features from different sources.

5. Acknowledgements

Each author contributed 50% equally to this work. The first author received scholarship support from CPALL for conducting this research in Panyapiwat Institute of Management.

References

[1] CCTV, 2015. *How Does Crowdfunding Not "Crowd-sorry"? Crowdfunding Failure Case Studies.* [online] Avaiable at: https://www.weiyangx.com/122711.html.

[2] Medium, 2019. *We Analyzed 331.000 Kickstarter Projects. Here´s What We Learned About Kickstarter Success.* [online] Avaiable at: https://medium.com/@daniel.kupka.

[3] Wang, W.Y., 2017. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, July 30-August 4, 2017, pp. 422-426.

[4] De Marneffe, M.C., Rafferty, A.N. and Manning, C.D., 2008. Finding contradictions in text. *Proceedings of ACL-08: HLT*, Columbus, Ohio, USA, June 19, 2008, pp. 1039-1047.

[5] Perez, B., Machado, S., Andrews, J. and Kourtellis, N., 2022. I Call BS: Fraud detection in crowdfunding campaigns. *14th ACM Web Science Conference*, Barcelona, Spain, June 26-29, 2022, pp. 1-11.

[6] Yu, Y.W. and Kim, H.G., 2020. Interactive morphological analysis to improve accuracy of keyword extraction based on cohesion scoring. *Journal of the Korea Society of Computer and Information*, 25(12), 145-153.

[7] Banawan, K. and Ulukus, S., 2018. The capacity of private information retrieval from coded databases. *IEEE Transactions on Information Theory*, 64(3), 1945-1956.

[8] Kim, S.W. and Gil, J.M., 2019. Research paper classification systems based on TF-IDF and LDA schemes. *Human-centric Computing and Information Sciences*, 9(1), 1-21.

[9] Yao, L., Pengzhou, Z. and Chi, Z., 2019. Research on news keyword extraction technology based on TF-IDF and TextRank. *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS),* Beijing, China, June 17-19, 2019, pp. 452-455.

[10] Li, W. and Zhao, J., 2016. TextRank algorithm by exploiting Wikipedia for short text keywords extraction. *2016 3rd International Conference on Information Science and Control Engineering*

(ICISCE), Beijing, China, July 8-10, 2016, pp. 683-686.

[11] Schmitt, X., Kubler, S., Robert, J., Papadakis, M. and LeTraon, Y., 2019. A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Granada, Spain, October 22-25, 2019, pp. 338-343.

[12] Contreras, J.O., Hilles, S. and Abubakar, Z.B., 2018. Automated essay scoring with ontology based on text mining and nltk tools. *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE),* Shah Alam, Malaysia, July 11-12, 2018, pp. 1-6.

[13] Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C. and Jatowt, A., 2020. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 509(1), 257-289.

[14] Campos, R., Mangaravite, V., Pasquali, A., Jorge, A.M., Nunes, C. and Jatowt, A., 2018. A text feature based automatic keyword extraction method for single documents. *European Conference on Information Retrieval*, Grenoble, France, March 26-29, 2018, pp. 684-691.

[15] Campos, R., Mangaravite, V., Pasquali, A., Jorge, A.M., Nunes, C. and Jatowt, A., 2018. Yake! collection-independent automatic keyword extractor. *European Conference on Information Retrieval*, Grenoble, France, March 26-29, 2018, pp. 806-810.

[16] Bowman, C.M., Danzig, P.B., Hardy, D.R., Manber, U. and Schwartz, M.F., 1995. The harvest information discovery and access system. *Computer Networks and ISDN Systems*, 28(1-2), 119-125.

[17] Gotz, D., When, Z., Lu, J., Kissa, P., Cao, N., Qian, W.H. and Zhou, M.X., 2010. Harvest: an intelligent visual analytic tool for the masses. *Proceedings of the First International Workshop on Intelligent Visual Interfaces for Text Analysis*, Hong Kong, China, February 7-10, 2010, pp. 1-4.

[18] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, USA, June 2-7, 2019, pp. 4171-4186.

[19] Sun, Y., Yang, D., Yu, T., Dong, A. and Yong, C., 2022. A study of BERT-based bi-directional Tibetan-Chinese neural machine translation. *International Conference on Computer, Artificial Intelligence, and Control Engineering (CAICE2022)*, Zhuhai, China, December 2, 2022, pp. 208-212.

[20] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A. and Raffel, C., 2021. mT5: A massively multilingual pre-trained text-to-text transformer. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Online, June 6-11, 2021, pp. 483-498.

[21] Qu, J., Nguyen, L.M. and Shimazu, A., 2016. Cross-language information extraction and auto evaluation for OOV term translations. *IEEE Transaction on China Communications*, 13(12), 277-296.

[22] Qu, J., Theeramunkong, T., Nguyen, L.M., Shimazu, A., Nattee, C. and Aimmanee, P., 2012. A flexible rule-based approach to Learn Medical English-Chinese OOV term translations from the web. *International Journal of Computer Processing of Languages*, 24(2), 207-236.

[23] Shu, K., Sliva, A., Wang, S., Tang, J. and Liu, H., 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), 22-36.

[24] Wenting, H. and Qu, J., 2022. Comparison of keyword extraction methods for crowdfunding projects based on web-data. *International Scientific Journal of Engineering and Technology*, 6(2), 1-12.

[25] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J., 2002. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* Philadelphia, USA., July 7-12, 2002, pp. 311-318.