Comparison of Keyword Etraction Methods for Crowdfunding Projects Based on Web-Data

Wenting Hou[1] and Jian Qu[2]

[1,2]Faculty of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi, Thailand
E-mail: 6372100134@stu.pim.ac.th, jianqu@pim.ac.th

Received: October 15, 2021 / Revised: December 17, 2021 / Accepted: December 17, 2021

***Abstract*—With the development of technology, there are more and more crowdfunding projects. However, it is hard for a human to understand such projects easily. Therefore, this study aims to provide a better solution for extracting keywords from each crowdfunding project so that everyone can quickly understand the core of these projects. In this study, we compared the performance of four keyword extraction methods on crowdfunding projects. The experimental results show that Bert performs better in precision, recall, and f-measure than NLTK, LIAAD, and Harvest algorithms. Moreover, we compared four pre-training models based on Bert and found that the distills-based-multilingual-cased-v1 model worked better than others with 74.0% in precision and 85.0% in F-measure.**

In addition, we also created a corpus of 106,869 pairs of text and its keyword for keyword extraction based on crowdfunding projects.

***Index Terms*—Crowdfunding Projects, Web-data, Searching API, Keyword Extraction, Bert Model**

I. Introduction

Why extract keywords?

In the era of the information explosion, information can be easily accessed on the internet, but humans cannot easily understand most of them. We need to extract some information that we are interested in. We can employ keyword extraction for such tasks. The extraction of keywords can also be called text label extraction. For example, "today's roast pork is really good", the word in the text "roast pork" can be considered a keyword or a label of this sentence. This keyword can express the meaning of the sentence to a certain extent. For example, if the word "roast pork" is used in a text classification task, it can imply information with the category of "food". There are normally two groups of methods for such keyword extraction: supervised and unsupervised methods. The supervised approach can achieve high accuracy, but the disadvantage is that it requires many labeled data and high labor costs. Compared with the supervised methods, the unsupervised methods have lower data requirements. Therefore, the application of such a method in the field of keyword extraction is more popular. In this study, we compare some of the common unsupervised keyword extraction algorithms, namely NLTK, Harvest, LIAAD, and a supervised method, namely Bert.

Why extract keywords from crowdfunding projects?

We found the failure rate of crowdfunding projects in 2015[1], 2017[2], 2019[3], 2020[4], and 2021[5], and we compared the failure rate of 2015, 2017, 2019, and 2021, as shown in Fig.1. From Fig.1 we can see that the failure rate of crowdfunding projects is high and has been rising except 2021.

[1]https://www.weiyangx.com/122711.html
[2]https://zhuanlan.zhihu.com/p/32325090
[3]https://medium.com/@daniel.kupka
[4]https://www.amz123.com/thread-348221.htm
[5]https://www.kickstarter.com/help/stats

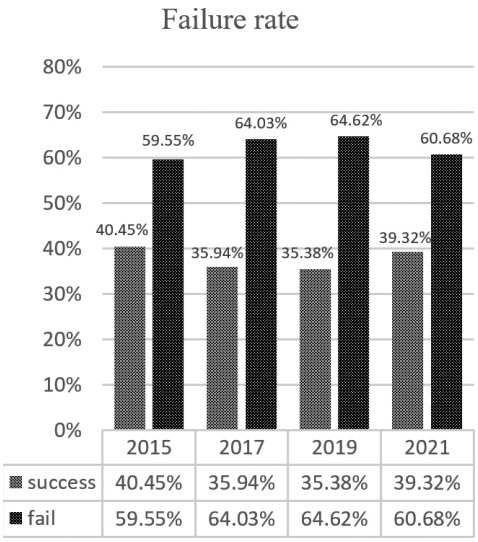

Fig. 1. Failure rate of crowdfunding projects

There are two reasons for the decline in the failure rate in 2021. Firstly, there were fewer crowdfunding projects this year, as shown in Table I. Although Kickstarter only displays all the data since its establishment in 2009, we can calculate that there were only 73,068 projects published in 2021 and there were 92518 projects published in 2020. There was a 21% reduction in the number of projects published. Secondly, the failure rate in 2021 was updated on December 2, 2021, missing a month. These two reasons lead to the decline in the failure rate in 2021.

TABLE I
COMPARISON OF NUMBER OF PROJECTS PUBLISHED IN 2020 AND 2021

List	2009-2019	2009-2020	2009-2021	2020	2021
Number of projects published	378000	470518	543586	92518	73068
Number of successful projects	133724	175169	212713	41445	37544

Since the high failure rate of crowdfunding projects, so we wanted to use keyword extraction to understand the main content of the project, which can be used to determine the feasibility of the project further. For example, the Fontus project claimed that it is a self-filling water bottle, especially designed to fit your bicycle. This water bottle will refill itself as you ride on your bike. It can create 0.5 liters of water per hour out of solar power and air. This device was designed to capture the moisture content in the air, condense it and store it as safe drinking water. The air stream you generate while riding is used here in place of a fan to pass large amounts of air into the chambers without needing extra energy sources. If you look at how the Fontus bottle advertises its ability to absorb water from the air, it seems logical to the layman. The Fontus bottle claims to be a small dehumidifier: it sucks air into the bottle and condenses to trap moisture when the air cools, using small solar panels to power the process. We all know how dehumidifiers work, so it seems simple enough. However, to produce 0.5 liters of water in 1 hour as Fontus claims, a 250-watt solar panel with a surface area of 1.5 square meters is required to operate at 100% efficiency! This is a huge, rooftop-sized solar panel--definitely not the small panel that Fontus is equipped with. It is less than 1/6 of the required panel size. Secondly, condensing moisture from humidity into water requires 9000 BTU/gal, if the air is already at the dewpoint, 100% humidity. 0.5 liters of water requires 1195 BTU to condense. If this water bottle was the efficient of a central air conditioner, this would use 92 watt-hrs of electricity. But Peltier solid-state cooling is like 10th the efficiency of air conditioning systems, so Fontus bottle needs about one kilowatt-hr. to make 0.5 liters of water in 100% humidity conditions. Although Fontus is equipped with a 250-watt solar panel, it would need four hours of work. This is a pretty huge contrast to the 0.5 liters per hour. In summary, this water bottle sounds logical but technically impossible.

Now, the product seems to have died, pitting many investors. It is because we do not know enough to see that Fonts does not have technical support to deliver the product. If we get the "self-filling water bottle" keyword extracted from the web data, we can quickly know the project. If we get the "produce 0.5 liters of water in 1 hour" keyword extracted from the web data, we can use it to search for feasibility on the Internet. Therefore, keyword extraction on crowdfunding projects is very important, due to such keywords might provide a better understanding of the project.

This is a relatively frontier research, we focus on keyword extraction and contradiction detection on crowdfunding projects. Most research on fake information is focused on fake news from social media, and research on fake crowdfunding projects is limited. So we decided to explore this direction, we thought that this innovative research will be more popular on crowdfunding platforms. This study is only the first step, the first step is the comparison of keyword extraction on crowdfunding projects, and the next step is the knowledge extraction to identify possible fake projects.

II. RELATED STUDIES

There are many methods for extracting keywords, roughly divided into the following three categories: statistical-based methods, graph-based methods, and semantic model-based methods. The statistical methods which we will explain include Term-Frequency (TF), Term Frequency-Inverse Document Frequency (TFIDF), Natural Language Toolkit (NLTK) [1] and

Yet Another Keyword Extractor (YAKE). Term-Frequency (TF) is the simplest method, which calculates the score by the frequency of words in the document. The main problem with word frequency is that it does not consider structural and semantic information and cannot distinguish synonyms. TFIDF is a simple but effective method proposed by Salton [2] in 1988, and it calculates each term in a text by considering two factors: one is the term's word frequency in the document, i.e., TF, and another is the Inverse Document Frequency (IDF), which measures how many texts contain the term. IDF is mainly used to penalize those terms that appear in many texts, and these terms are usually some irrelevant deactivation words, etc. The whole core idea of TFIDF [3] is that the importance of a term in a document depends on the frequency of the term in the document and the number of occurrences in other documents. However, the TFIDF algorithm also has obvious drawbacks. It is not comprehensive enough to measure the importance of a word by its frequency simply, and sometimes the important words may not appear many times [4]. NLTK is a well-known natural language processing library for Python [5], which comes with classification, word separation and other functions. In this study, we combined RAKE and NLTK to form RAKE-NLTK, it achieved 62.4% of F-measure based on the GMB corpora [1]. If applied to crowdfunding project, the results might be different. Harvest Text [6] is a library that focuses on unsupervised (weak) methods and can integrate domain knowledge (e.g., types, aliases) for simple and efficient processing and analysis of a domain-specific text. It has many features such as text cleaning, new word discovery, sentiment analysis, entity recognition linking, keyword extraction, knowledge extraction, syntactic analysis, etc. David Gotz et al. [7] presented an intelligent visual analytic system called Harvest, which was designed to empower everyday business users to derive insight from large amounts of data. They found that there was a 75% reduction in error rate on average. When a task was performed using Many Eyes, it achieved 22% of error rate; when a task was performed using HARVEST, it achieved 5.6% of error rate. They attributed the sharp drop in error rate to Harvest's ability that can let users easily explore data from different angles. Yet Another Keyword Extractor (YAKE) is an unsupervised keyword extraction algorithm[8]. It relies on statistical features of text extracted from a single document to select the most important keywords in the text. Ricardo Campos et al. [9] proposed YAKE to extract keywords from single documents, and compared it with RAKE, TextRank, SingleRank and TFIDF. Based on the Schutz2008 database, YAKE achieved 9.1% of F-measure, TextRank achieved 8.2% of F-measure, TFIDF achieved 4.3% of F-measure, SingleRank achieved 3.7% of F-measure, RAKE achieved 0.6% of F-measure. So, YAKE performed better. The core idea of the statistical-based approach is to calculate the score of each word or phrase in the text, and it is possible that all words can be sorted with the scores, then the top n words with the highest scores are obtained as the keywords of the text [10]. The statistical features include co-occurrence frequency, symmetric conditional probability, modified association measure, chi-square, mean distance, length similarity, and word frequency. In medical or biological fields, many information extraction systems and studies rely on a certain corpus. Qu et al. [11] proposed a new approach to address English medical OOV terms. Unlike most existing methods for translating English OOV terms into Chinese, their candidates are selected by a machine learning system with the support of different features, and the best candidate selection results in the highest correct rate of 86.79% using features such as lift, frequency, and distance together. This suggests we may employ more features to find a better keyword.

Secondly, the Graph-Based Approaches include PageRank [12], TextRank, SingleRank, TopicRank, and PositionRank. PageRank was first used to calculate the importance of web pages, and TextRank is a graph-based ranking algorithm for text [13]. The basic idea is derived from Google's Page Rank algorithm, which automatically extracts many meaningful words or phrases from a given text. The original text is split into sentences. In each sentence, deactivated words are filtered out, and only words of the specified lexical nature are retained. It results in a collection of sentences and a collection of words [14]. TopicRank treats topics as clusters of similar key phrases [15], which are ranked according to their importance in the document, then top n most relevant topics are selected, and each topic selects one most important key phrase to represent the core keywords of the document.

Thirdly, Semantic Models include Linear Discriminant Analysis (LDA) [16], Hidden Markov Model (HMM) [17], Recurrent Neural Networks (RNN), and Bidirectional Encoder Representations from Transformers (Bert), etc. The keyword or phrase extraction based on semantic models is generally supervised learning. It treats keyword extraction as an annotation task to determine whether the word is a keyword or not; or after classifying the text, it automatically learns the weight score of each word in the text based on the attention layer, and extracts keywords according to the score. These methods are all supervised learning and require labeled data for training the model. Zhang et al. [18] proposed a 2-layer RNN model that treats keyword and key phrase extraction as an annotation classification task to determine whether each word is a keyword or a key phrase. The first layer of the model is used for the keyword recognition task, and the second layer is

used for the key phrase recognition task. Fusion weights the loss functions of two tasks as the final loss function. The proposed RNN model achieved 80.97% of F-measure on automatically extracting keyphrases from single tweets. BERT is a pre-trained language model, and the full name is Bidirectional Encoder Representations from Transformer. It means that it is a bidirectional encoder representation based on Transformer [19]. As the name suggests, Bert uses the Transformer. It can take the word that precedes and follows it into account while processing a word to get its meaning in the context. We know that Transformer's attention mechanism has a good effect on feature extraction of words in context. Overall, the Bert model uses the popular feature extractor Transformer and implements a bi-directional language model, giving it good performance. Yili Qian et al. [20] proposed a text keyword extraction method based on Bert, and compared it with TFIDF, TextRank, and LDA. Based on 300 scientific papers downloaded from Wanfang database, TFIDF achieved 36.4% of F-measure, TextRank achieved 40.7% of F-measure, LDA achieved 42.0% of F-measure, a keyword extraction algorithm based on Bert and multi class feature fusion achieved 43.6% of F-measure. The results show that the combination algorithm based on Bert is better than the single extraction algorithm. However, there is still room for improvement. For example, Bert uses the original Transformer. Although it is powerful, now there are some more powerful and improved versions; another place left to be improved is the Mask mechanism of Bert, which can be used to train the Bidirectional language model, but this will lead to inconsistency between pre-training and fine-tuning on downstream tasks.

Fig. 2. The overall flow chart of the study

input	url	title	summary
"Star Citizen"	robertsspaceindustries.com/	...Space Industries | Follow the development of Star C...	查看此网页的中文翻译，请点击 翻译此页 Roberts Space Industries is the official go-to website for all news about Star Citizen and Squadron 42. It also hosts the online store for game items...
	starcitizen.mmmos.com/	Home Page : Star Citizen	查看此网页的中文翻译，请点击 翻译此页 Star Citizen is a space trading and combat sim. It will run on windows and linux. It was 100% crowd funded on
	starcitizen.howar31.com/	首頁| Star Citizen - 星際公民中文社群網	這裡是 Star Citizen 星際公民 中文社群網, 在這裡將會提供星際入門指南、民生與軍用品介紹、船艦交通工具展示、UEE地球聯合帝國最新消
	fanyi.baidu.com	star citizen - 百度翻译	star citizen 英[stɑ:(r) ˈsɪtɪzn] 美[stɑːr ˈsɪtɪzn] 网络 星际公民; [例句]But as of early November, the focus had shifted to issues such as the move by Gong Li, the film star, to become a Singaporean citizen. 但到11月初，焦点转移到了影星成为新加坡公民之类的问题上。
	豆瓣	星际公民 Star Citizen (豆瓣)	2019年2月16日 Star Citizen 星际公民的视频, 攻略, 评测, 图片, 评分, 讨论, 帮助你判断是否好玩, 发现更多相似好游戏及爱玩这些游戏的
	starcitizen.jeuxonline.info/	Star Citizen	Star Citizen : mise à jour Alpha 3.10 et accès gratuit jusqu'au 23 septembre Star Citizen 12 septembre 42 Après
	www.yxbao.com/game/558...html	星际公民Star Citizen中文版下载_星际公民Star	2017年5月20日 《StarCitizen中文版》(星际公民)是一款模拟经营游戏, 它除了一款PC平台的星际游戏之外, 它更是一个深化, 一个众筹神话。
	starcitizen.wikia.com/	Star Citizen Wiki | Fandom	Agents of Mayhem • Battalion 1944 • Battleborn • Battlefield • Borderlands • Brothers in Arms • Bulletstorm
	知乎	如何评价太空游戏新作《星际公民》(Star	2018年4月20日 而这就是《Star Citizen》最大的问题:大把的精力花在飞船设计上了。做几个土豪版飞船看得见摸得着, 群众喜闻乐见, 还能
	www.kickstarter.com/projects/c...	Star Citizen by Cloud Imperium Games	Cloud Imperium Games Corporation 正在 Kickstarter 上為 Star Citizen 籌款! Reclaim the stars in the exciting new Space

Fig. 3. Example of snippet retrieval

III. Our Approach

In this section, we describe our approach, and it has two major steps: information retrieval and keyword extraction, as shown in Fig. 2. Next, we introduce the information retrieval and the keyword extraction.

A. Information Retrieval

We found a list of the highest-funded crowdfunding projects on Wikipedia[1]. With the data from Wikipedia, we used the Internet to retrieve information. The first step is snippet retrieval of the project. We separated the retrieved snippet into three different fields in the database: URL, title, and summary. Moreover, input is the project's name, as shown in Fig. 3.

The name of the crowdfunding projects may be ambiguous, thus the retrieved information may not be related to the project. For example, we search for the "EOS" project, and there may be a person or a song called "EOS". Therefore, we need to find an efficient way for making our query text less ambiguous. There are four ways of constructing the query text: searching the name of the project, searching the name of the project and its category, searching the name of the project and its crowdfunding platform, and searching the name of the project plus category plus crowdfunding platform, as shown in Table II. We compare A to the name of the project, B to the category of the project, C to the crowdfunding platform of the project.

TABLE II
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

We conducted a test to find out which search method is better. We searched the first 20 projects of the list of highest-funded crowdfunding projects on Wikipedia. Each project intercepted ten snippets, 20 projects intercepted 200 snippets. The feedback snippets of each search method are different, So we got 800 different snippets in total. We used 800 snippets to compare the effective ratios in these four search methods. We judged the correctness of each snippet. When the input is "A", the number of correct snippets is 143, the effective ratio is 71.5%; And when the input is "A+B", the number of correct snippets is 180, the effective ratio is 90%; When the input is "A+C", the number of correct snippets is 115, the effective ratio is 57.5%; And when the input is "A+B+C", the number of correct snippets is 129, the effective ratio is 64.5%. By calculating 800 snippets, we found that searching the name of the project and its category to retrieve snippets is the most efficient and least intrusive way of constructing the query text. The effective ratio is arranged from high to low, as shown in Table III.

TABLE III
Effective Ratios of Different Input of Snippet Retrieval

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%

In summary, this study selected the input "A+B" for Internet snippet retrieval. We set up a programmable search engine by calling the API. We changed the region, language, and website in the basic setting of the programmable search engine. The region is set to all regions, which returns more contents; the language is set to simplified Chinese, because this study is more concerned with extracting Chinese characters. The websites to be searched are shown in Table IV. These ten websites often appeared when we searched project information manually.

TABLE IV
List of Websites

Number	Website
1	http://www.doc88.com
2	sogou.com
3	weibo.com
4	www.zhihu
5	www.bing.com
6	www.csdn.net
7	www.sohu.com
8	https://zol.com.cn
9	www.baidu.com
10	www.google.com

Then we used MySQL to create a database and created a table containing ID, Keyword, URL, Title, Summary. Next, we employed Google Search API to retrieve information automatically, and such information was saved into the database.

[1]https://en.wikipedia.org/wiki/List_of_highest-funded_crowdfunding_projects (last accessed on 1 October 2021)

B. Keyword Extraction

We reviewed papers to find seven best keyword extraction methods. Of these seven methods, only four perform well in Chinese. These four methods are NLTK, LIAAD, Harvest, and Bert. Among these four methods, there is only one supervised method called Bert. So we make up for this shortcoming by comparing four models based on Bert.

Firstly, we compared NLTK, LIAAD, Harvest, and Bert using three different types of documents. Although we limited the language to simplified Chinese, there were still English results. “testen” is a pure English document; “testch” is a pure Chinese document; the “test” document is half Chinese and half English. The results show that the Bert model is better, and the selected candidates are closer to the document’s meaning, as shown in Table V.

TABLE V
DIFFERENT METHODS FOR EXTRACTING KEYWORDS

	Testen	Testch	Test
Original text	The iBackPack has the capability to hold all of your electronics. There is an optional built-in WiFi connection and batteries galore. It includes a 20,000 mAh primary, 8,000 mAh secondary, and a half a dozen other batteries that ensure your electronics are constantly charged.	兵马俑，即秦始皇兵马俑，亦简称秦兵马俑或秦俑，第一批全国重点文物保护单位，第一批中国世界遗产，位于今陕西省西安市临潼区。兵马俑是古代墓葬雕塑的一个类别。古代实行人殉，奴隶是奴隶主生前的附属品，奴隶主死后奴隶要作为殉葬品为奴隶主陪葬。兵马俑即制成兵马（战车、战马、士兵）形状的殉葬品 。	基于广受赞誉的 Stellaris PC 游戏，Stellaris Infinite Legacy 提供了您喜欢的 4x 棋盘游戏，其中包括个性化的定制内容和突如其来的故事，使 Stellaris PC 游戏与众不同。Stellaris Infinite Legacy 是一款可供 2 至 4 名玩家使用的 2 小时 4x 棋盘游戏，其简单规则会根据您在游戏中的选择而增长
NLTK	‘000 mah secondary’, ‘000 mah primary’, ‘wifi connection’, ‘optional built’, ‘constantly charged’	‘士兵（形状的殉葬品’, ‘兵马俑即制成兵马）战车’, ‘第一批全国重点文物保护单位’, ‘第一批中国世界遗产’, ‘战马’	‘stellaris infinite legacy 是一款可供 2 至 4 名玩家使用的 2 小时 4x 棋盘游戏’, ‘stellaris infinite legacy 提供了您喜欢的 4x 棋盘游戏’, ‘基于广受赞誉的 stellaris pc 游戏’, ‘使 stellaris pc 游戏与众不同’, ‘其简单规则会根据您在游戏中的选择而增长’
Harvest	/	‘兵马俑’, ‘奴隶主’,’ 殉葬品’, ‘奴隶’,’ 临潼区’	‘游戏’, ‘棋盘’, ‘赞誉’, ‘玩家’, ‘个性化’ ‘基于广受赞誉的 stellaris’,
Bert	‘dozen’, ‘electronics’, ‘batteries’, ‘20’, ‘wifi’	‘第一批中国世界遗产’, ‘第一批全国重点文物保护单位’, ‘奴隶主死后奴隶要作为殉葬品为奴隶主陪葬’, ‘兵马俑是古代墓葬雕塑的一个类别’, ‘奴隶是奴隶主生前的附属品’	‘其简单规则会根据您在游戏中的选择而增长’, ‘其中包括个性化的定制内容和突如其来的故事’, ‘legacy 提供了您喜欢的 4x 棋盘游戏’, ‘legacy 是一款可供 2 至 4 名玩家使用的 2 小时 4x 棋盘游戏’
LIAAD	‘the’, ‘your’, ‘electronics’, ‘ibackpack’, ‘has’	‘兵马俑，即秦始皇兵马俑，亦简称秦兵马俑或秦俑，第一批全国重点文物保护单位，第一批中国世界遗产，位于今陕西省西安市临潼区。兵马俑是古代墓葬雕塑的一个类别。古代实行人殉，奴隶是奴隶主生前的附属品，奴隶主死后奴隶要作为殉葬品为奴隶主陪葬。兵马俑即制成兵马（战车、战马、士兵）形状的殉葬品’	‘stellaris’, ‘infinite’, ‘legacy’, ‘名玩家使用的’, ‘棋盘游戏，其简单规则会根据您在游戏中的选择而增长’,

However, individual cases do not represent the whole, and we will use more data to determine whether the Bert model is the most suitable method for this study. Next, we will explain to you the four methods we compared.

1) NLTK

NLTK is a natural language processing library for Python. A virtual example is shown in Table VI.

TABLE VI
A VIRTUAL EXAMPLE OF NLTK

Original text	兵马俑，即秦始皇兵马俑，亦简称秦兵马俑或秦俑，是第一批全国重点文物保护单位
Input	C1C2C3P1C4C5C6C7C8C9C10P2C11C12C13C14 C15C16C17C18C19C20P3C21C22C23C24C25C26 C27C28C29C30C31C32C33C34
Output	C21C22C23C24C25C26C27C28C29C30C31C32 C33C34, C4C5C6C7C8C9C10
Answer	是第一批全国重点文物保护单位 即秦始皇兵马俑

2) LIAAD

Yake is a lightweight unsupervised automatic keyword extraction method mentioned. A virtual example is shown in Table VII.

TABLE VII
A VIRTUAL EXAMPLE OF LIAAD

Original Text	兵马俑，即秦始皇兵马俑，亦简称秦兵马俑或秦俑，是第一批全国重点文物保护单位
Input	C1C2C3P1C4C5C6C7C8C9C10P2C11C12C13C14 C15C16C17C18C19C20P3C21C22C23C24C25C26 C27C28C29C30C31C32C33C34
Output	C1C2C3P1C4C5C6C7C8C9C10P2C11C12C13C14 C15C16C17C18C19C20P3C21C22C23C24C25C26 C27C28C29C30C31C32C33C34
Answer	兵马俑，即秦始皇兵马俑，亦简称秦兵马俑或秦俑，是第一批全国重点文物保护单位

3) Harvest

HarvestText has many features such as keyword extraction, knowledge extraction, etc. In this study, we used it to obtain keywords in the text based on algorithms such as Textrank, tfidf, *etc.*, using JIEBA for word separation and TFIDF for extraction. A virtual example is shown in Table VIII.

TABLE VIII
A VIRTUAL EXAMPLE OF HARVEST

Original Text	兵马俑，即秦始皇兵马俑，亦简称秦兵马俑或秦俑，是第一批全国重点文物保护单位
Input	C1C2C3P1C4C5C6C7C8C9C10P2C11C12C13C14 C15C16C17C18C19C20P3C21C22C23C24C25C26 C27C28C29C30C31C32C33C34
Output	C1C2C3, C14C15C16C17
Answer	兵马俑 秦兵马俑

4) Bert

BERT is a pre-trained language model mentioned. In this study, we used it for keyword extraction. A virtual example is shown in Table IX.

TABLE IX
A VIRTUAL EXAMPLE OF BERT

Original Text	兵马俑，即秦始皇兵马俑，亦简称秦兵马俑或秦俑，是第一批全国重点文物保护单位
Input	C1C2C3P1C4C5C6C7C8C9C10P2C11C12 C13C14C15C16C17C18C19C20P3C21C22C23C24 C25C26C27C28C29C30C31C32C33C34
Output	C21C22C23C24C25C26C27C28C29C30C31C32 C33C34, C1C2C3
Answer	是第一批全国重点文物保护单位 兵马俑

Secondly, we compared four pre-trained models based on Bert for keyword extraction. Assuming that the most similar candidate to the document is a good keyword/keyphrase representing the document, converting the document and the candidate into a vector, we used the cosine similarity between the vectors to calculate the similarity between the candidate and the document. The top five most similar candidates of the document are used as the resultant keywords, as shown in the Table X. M1=quora-distilbert-multilingual, M2=distilbert-base-nli-mean-tokens, M3=distiluse-base-multilingual-cased-v1, M4=distiluse-base-multilingual-cased-v2. After comparison, we found that the M2 worked better, showing excellent performance in the similarity task. We applied the comparison to the large-scale data in an attempt to draw a conclusion that was not individual cases. We searched all 120 projects to generate 5340 snippets. For these 5340 snippets, we used four models based on Bert to generate 64,044 keywords, such information was saved into the database, as shown in Fig. 4. These 64,044 keywords are the large-scale data for further research.

TABLE X
KEYWORD EXTRACTION UNDER DIFFERENT MODELS BASED ON BERT

Model	Top 5	Characteristic
Human results	‘兵马俑’, ‘亦简称秦兵马俑或秦俑’, ‘第一批中国世界遗产’, ‘第一批全国重点文物保护单位’, ‘兵马俑是古代墓葬雕塑的一个类别’	
M1	‘兵马俑即制成兵马’, ‘兵马俑是古代墓葬雕塑的一个类别’, ‘形状的殉葬品’, ‘古代实行人殉’, ‘奴隶主死后奴隶要作为殉葬品为奴隶主陪葬’	Use parallel data in more than 50 languages and fine-tune
M2	‘第一批中国世界遗产’, ‘第一批全国重点文物保护单位’, ‘奴隶主死后奴隶要作为殉葬品为奴隶主陪葬’, ‘兵马俑是古代墓葬雕塑的一个类别’, ‘奴隶是奴隶主生前的附属品’	STSb performance: 85.16

TABLE X
KEYWORD EXTRACTION UNDER DIFFERENT MODELS BASED ON BERT (CON.)

Model	Top 5	Characteristic
M3	‘亦简称秦兵马俑或秦俑’, ‘兵马俑’, ‘兵马俑即制成兵马’, ‘即秦始皇兵马俑’, ‘兵马俑是古代墓葬雕塑的一个类别’	Support 15 languages
M4	‘第一批中国世界遗产’, ‘奴隶是奴隶主生前的附属品’, ‘亦简称秦兵马俑或秦俑’, ‘奴隶主死后奴隶要作为殉葬品为奴隶主陪葬’, ‘兵马俑是古代墓葬雕塑的一个类别’	Support more than 50 languages

input	summaryID	summary	kw1	kw2	kw3	kw4
Star Citizen + 电子游戏	1	('You can download and play Star Citizen Alpha 3....	电子游戏	download	alpha	alpha
Star Citizen + 电子游戏	1	('You can download and play Star Citizen Alpha 3....	会公示有助于你了解公共主页用途的信息	会公示有助于你了解公共主页用途的信息	公共主页信息公示查看更多	公共主页信息公示查看更多
Star Citizen + 电子游戏	1	('You can download and play Star Citizen Alpha 3....	download	facebook	会公示有助于你了解公共主页用途的信息	会公示有助于你了解公共主页用途的信息
Star Citizen + 电子游戏	2	('RSI's Spectrum is our integrated community and p...	rsi	integration	community	community
Star Citizen + 电子游戏	2	('RSI's Spectrum is our integrated community and p...	game	interaction	integrated	integrated
Star Citizen + 电子游戏	2	('RSI's Spectrum is our integrated community and p...	spectrum	forums	integration	integration
Star Citizen + 电子游戏	3	('电子游戏. Satisfactory. 电子游戏. 这个公共主页赞了. IGN. 公共主页发布的...	这个公共主页赞了	这个公共主页赞了	这个公共主页赞了	电子游戏
Star Citizen + 电子游戏	3	('电子游戏. Satisfactory. 电子游戏. 这个公共主页赞了. IGN. 公共主页发布的...	4月30日20	公共主页发布的近期帖子	公共主页发布的近期帖子	war
Star Citizen + 电子游戏	3	('电子游戏. Satisfactory. 电子游戏. 这个公共主页赞了. IGN. 公共主页发布的...	公共主页发布的近期帖子	4月30日20	电子游戏	公共主页发布的近期帖子
Star Citizen + 电子游戏	4	('《星际公民》（英文：Star Citizen）是已经在Microsoft Windows和Lin...	星际公民	也有正在开发中的单人	星际公民	星际公民
Star Citizen + 电子游戏	4	('《星际公民》（英文：Star Citizen）是已经在Microsoft Windows和Lin...	windows和linux公开的太空模拟电子游戏	星际公民是一个mmorpg游戏	windows和linux公开的太空模拟电子游戏	星际公民是一个mmorpg游戏
Star Citizen + 电子游戏	4	('《星际公民》（英文：Star Citizen）是已经在Microsoft Windows和Lin...	星际公民是一个mmorpg游戏	windows和linux公开的太空模拟电子游戏	星际公民是一个mmorpg游戏	windows和linux公开的太空模拟电子游戏

Fig. 4. Example of different keywords extracted by four models based on Bert in database

IV. RESULTS AND DISCUSSION

In this section, as shown in Fig.5, we describe the results of information retrieval and keyword extraction.

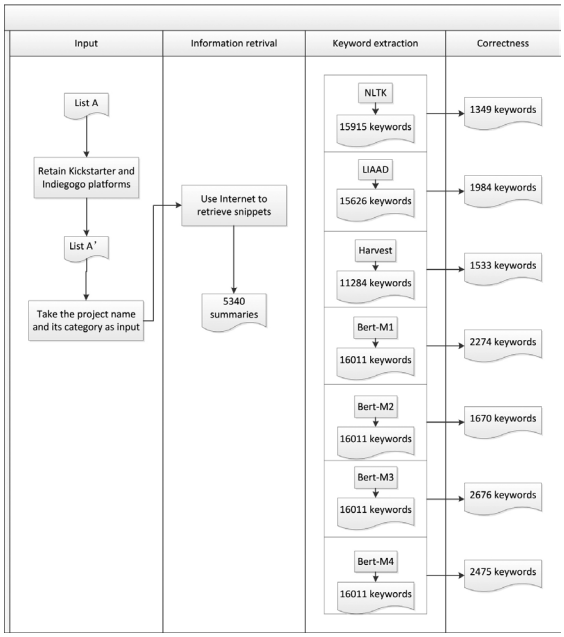

Fig. 5. The overall flow chart of the project with results

A. Data Source

This study used two datasets: List A, List A’. List A is the original dataset on Wikipedia, which is a list of the highest-funded crowdfunding projects (including those that failed to receive funding), and List A’ is a list of projects from A with only Kickstarter and IndieGoGo platforms retained. In this study, we focus on the keyword extraction of the 5340 summaries retrieved from 120 projects in List A’ to select a better method and model.

For these 5340 snippets, we used NLTK, LIAAD, and Harvest to generate 42,825 keywords, and used four models based on Bert to generate another set of 64,044 keywords. We hired five master students to help us mark the correctness of these keywords. Finally, we created a corpus to storage those tagged 106,869 pairs of text and its keyword for keyword extraction based on crowdfunding projects.

B. Method Comparison

After the keyword extraction, the next step is evaluation. The project-related keywords will be selected from all the keywords extracted. Different methods resulted in 58836 keywords. We artificially marked correctness to derive the precision, recall, and F-measure of each method in the field of crowdfunding projects.

TABLE XI
TOTAL NUMBER OF KEYWORDS IN DIFFERENT METHODS

Method	Total Number of Keywords	Correctness
NLTK	15915	1349
LIAAD	15626	1984
Harvest	11284	1533
Bert	16011	2676

There are 120 projects. Each project retrieves 100 summaries, sometimes less than 100, and ends up with 5340. We took 5340 summaries as input, and set the number of keywords extracted to 3 for each method. Thus, NLTK got 15915 keywords, LIAAD got 15626 keywords, Harvest got 11284 keywords, and Bert got 16011 keywords. As shown in Table XI, Harvest extracts the least keywords because it is based on the jieba-tfidf algorithm for keyword extraction, while JIEBA is only applicable to only Chinese word separation[1], so the difference between the Harvest and other methods is as high as four thousand, but the content of this study focuses on Chinese content, so we only consider the precision rate, recall rate, and F-measure.

As shown in Table XII, the results show that the Bert model works better because the selected candidates are closer to the document's meaning.

TABLE XII
COMPARISON OF DIFFERENT METHODS

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%

From Fig. 6 we can see that NLTK has the lowest F-measure, only 15.6%. Bert has the highest F-measure, with 28.6%. It is better than the other three methods, the most important reason is that only Bert can obtain the bidirectional feature representation of the context among these four keyword extraction methods. So, Bert is the most suitable method for extracting keywords on crowdfunding projects.

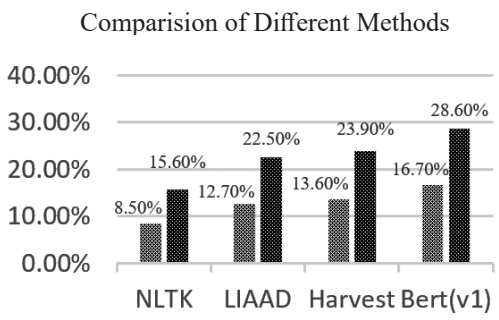

Fig. 6. Comparison of different methods

C. Model Comparison

Next, the comparison of different models based on Bert is shown in Table XIII. Different models resulted in 64044 keywords, we marked the correctness of keywords and calculated the F-measure of each model, as shown in Table XIII. M1= quora-distilbert-multilingual, M2=distilbert-base-nli-mean-tokens, M3=distiluse-base-multilingual-cased-v1, M4=distiluse-base-multilingual-cased-v2. And "0" means incorrect, "1" means correct. From Table XIII, we can see that M3 works better.

TABLE XIII
COMPARISON OF DIFFERENT MODELS WITH 64044 KEYWORDS

	Total Number of Keywords	1	0	Precision	Recall	F-Measure
M1	16011	2274	13737	14.2%	1	24.9%
M2	16011	1670	14341	10.4%	1	18.8%
M3	16011	2676	13335	16.7%	1	28.6%
M4	16011	2475	13536	15.5%	1	26.8%

From Fig.7 we can see that M3 has the highest F-measure of 28.6%. While M2 only has 18.8% of F-measure. Although an example shows that M2 performs better, but large-scale data shows that M3 performs well.

[1]https://github.com/fxsjy/jieba (last accessed on 1 October 2021)

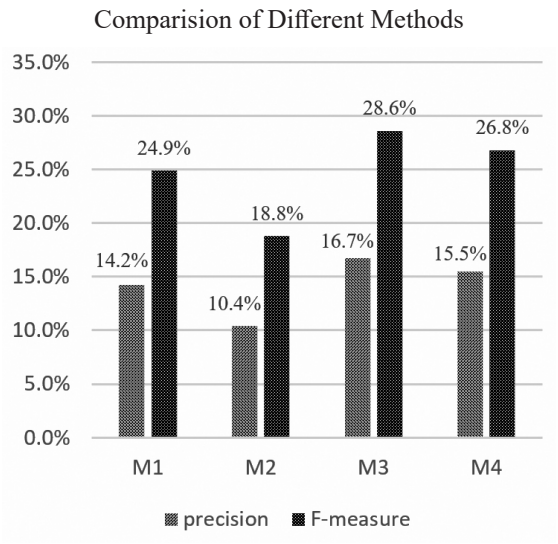

Fig. 7. Comparison of different models

In order to find out the effect of data size, we divided the number of snippets feedback into four sub-sets: TOP25, TOP50, TOP75 and TOP100. We compared the performance of Bert-M1, Bert-M2, Bert-M3 and Bert-M4 under these four sub-sets. When only the first 25 snippets were taken for each project, M3 achieved the highest F-measure of 30.0%; When the first 50 snippets were taken for each project, M3 achieved the highest F-measure of 30.1%; When the first 75 snippets were taken for each project, M3 achieved the highest F-measure of 29.1%; When the first 100 snippets were taken for each project, M3 achieved the highest F-measure of 28.6%. We found that the performance of M3 is always the best model regardless of the size of the data, as shown in Table XIV.

TABLE XIV
COMPARISON OF DIFFERENT METHODS ON DIFFERENT SIZES OF DATA

Method	TOP N	TOP25	TOP50	TOP75	TOP100
	True	881	1293	1340	1349
	Total	8449	13625	15348	15915
NLTK	Precision	10.4%	9.5%	8.7%	8.5%
	Recall	65.3%	95.8%	99.3%	100%
	F-score	**18.0%**	**17.3%**	**16.1%**	**15.6%**
	True	1283	1877	1974	1984
	Total	8382	13395	15062	15626
LIAAD	Precision	15.3%	14.0%	13.1%	12.7%
	Recall	64.7%	94.6%	99.5%	100%
	F-score	**24.8%**	**24.4%**	**23.2%**	**22.5%**
	True	967	1443	1523	1533
	Total	6401	10125	11070	11284
Harvest	Precision	15.1%	14.3%	13.8%	13.6%
	Recall	63.1%	94.1%	99.3%	100%
	F-score	**24.4%**	**24.8%**	**24.2%**	**23.9%**
	True	1461	2121	2251	2274
	Total	8500	13717	15444	16011
Bert-M1	Precision	17.2%	15.5%	14.6%	14.2%
	Recall	64.2%	93.3%	99.0%	100%
	F-score	**27.1%**	**26.6%**	**25.4%**	**24.9%**
	True	1076	1551	1649	1670
	Total	8500	13717	15444	16011
Bert-M2	Precision	12.7%	11.3%	10.7%	10.4%
	Recall	64.4%	92.9%	98.7%	100%
	F-score	**21.2%**	**20.1%**	**19.3%**	**18.8%**
	True	1674	2465	2634	2676
	Total	8500	13717	15444	16011
Bert-M3	Precision	19.7%	18.0%	17.1%	16.7%
	Recall	62.6%	92.1%	98.4%	100%
	F-score	**30.0%**	**30.1%**	**29.1%**	**28.6%**
	True	1550	2277	2439	2475
	Total	8500	13717	15444	16011
Bert-M4	Precision	18.2%	16.6%	15.8%	15.5%
	Recall	62.6%	92.0%	98.5%	100%
	F-score	**28.2%**	**28.1%**	**27.2%**	**26.8%**

We also compared the performance of NLTK, LIAAD and Harvest under these four sub-sets, as shown in Table XIV. From the data point of view, the smaller the TOPN, the smaller the recall rate. However, as TOPN becomes larger, the accuracy rate will also decrease under normal circumstances. So the performance can be judged by the F-measure in combination. We found TOP25 always had the highest F-measure, because the noise was minimal at this time. We can think that the higher the ranking of the snippets retrieved on the Internet, the more relevant the snippet and the project, and the higher the correct keyword extraction rate. But there are two exceptions: when the method is Bert-M3, TOP50 has the highest F-measure of 30.1%; and when the method is Harvest, TOP50 has the highest F-measure of 24.8%. These two methods have better anti-noise performance, because they change little with the change of TOPN.

In summary, the experimental results show that the M3 model performs best When the first 50 snippets are taken for each project.

In addition, in order to show that when the first 50 snippets are taken for each project, Bert's M3 model is the most suitable method for this study, we compared the keywords extracted by Bert with the keywords extracted by human. We computed the ratio of the edit distance to the length of max (string1, string 2). 0 means that the sequences are identical, while 1.0 means that they have nothing in common. When the ratio of the edit distance is between 0-0.6, we think that the keywords extracted by Bert are true. We used the keywords extracted manually as the ground truth, and found that F-measure was as high as 74.0%, F-measure was as high as 85.0%. The details are shown in Table XV.

TABLE XV
COMPARISON OF KEYWORDS EXTRACTED BY MACHINE AND HUMAN

Human Results	M3	M3-Edit Distance	Precision	Recall	F-Measure
3192	2465	1823	74.0%	100%	85.0%

V. CONCLUSION AND FUTURE WORK

In this research, we proposed Bert to extract keywords from crowdfunding projects, and compared it with NLTK, LIAAD and Harvest, Bert performed best. Compared with the four models based on Bert, M3 performed best. Based on 106,869 pairs of keywords, Bert's M3 model is the best keyword extraction method for crowdfunding projects. And when retrieving TOP50 snippets, M3 performed better, it achieved 85.0% of F-measure. Keyword extraction is widely used in the field of NLP. If we can accurately describe the document with a few simple keywords, we can understand whether an article is what we need by just looking at a few keywords, which will greatly improve our information acquisition efficiency.

In the future, we plan to study the effect of mixing these Bert models on the keyword extraction of crowdfunding projects. Because the five candidates selected by M3 are not all optimal in TABLE IX: it does not propose the word "第一批中国世界遗产", while M2 does. So, we may consider a mixture of several models. For example, we may use the M3 to select the first two candidates and use the M2 to select the first three candidates, so that the combined five candidates will be more similar to the text than the candidates selected by a single model. We think selecting candidates by a mixture of different models will be more similar to the keywords that were selected by human.

ACKNOWLEDGEMENTS

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 second authors contribute 50% equally to this work.

REFERENCES

[1] X. Schmitt, S. Kubler, J. Robert et al., "A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate," in *Proc. 2019 Sixth International Conference on Social Networks Analysis, Management and Security*, 2019, pp. 338-343.

[2] G. B. Salton and C. Buckley, "Term-Weighting Approaches In Automatic Text Retrieval," *Information Processing and Management*, vol. 24, no. 5, pp. 513-523, Jan. 1988.

[3] T. Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization," in *Proc. The 14th International Conference on Machine Learning,* 1997, pp. 143-151.

[4] L. Huang, Y. Wu, and Q. J. C. S. Zhu, "Research and Improvement of Keyword Automatic Extraction Method," *Journal of Computer Science*, vol. 41, no. 6, pp. 204-207, Jun. 2014.

[5] E. Loper and S. Bird, "NLTK: The Natural Language Toolkit," in *Proc. The COLING/ACL 2006 Interactive Presentation Sessions*, 2006, pp. 69-72.

[6] C. M. Bowman, P. B. Danzig, D. R. Hardy et al., "The harvest Information Discovery and Access System," *Computer Networks and ISDN Systems*, vol. 28, no. 1-2, pp. 119-125, Dec. 1995.

[7] D. Gotz, Z. When, J. Lu et al., "Harvest: an Intelligent Visual Analytic Tool for the Masses," in *Proc. The First International Workshop on Intelligent Visual Interfaces for Text Analysis*, 2010, pp. 1-4.

[8] 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, Jan. 2020.

[9] R. Campos, V. Mangaravite, A. Pasquali et al., "A Text Feature Based Automatic Keyword Extraction Method for Single Documents," in *Proc. European Conference on Information Retrieval*, 2018, pp. 684-691.

[10] R. Campos, V. Mangaravite, A. Pasquali et al., "Yake! Collection-Independent Automatic Keyword Extractor," in *Proc. European Conference on Information Retrieval*, 2018, pp. 806-810.

[11] J. Qu, T. Theeramunkong, C. Nattee et al., "Web Translation of English Medical OOV Terms to Chinese with Data Mining Approach," *Science and Technology Asia*, vol. 16, no. 2, pp. 26-40, Jun. 2011.

[12] L. Page, S. Brin, R. Motwani et al., (1998, Jan. 28). *The PageRank Citation Ranking: Bringing Order to the Web*. [Online]. Available: http;//ilpubs.stanford.edu:090/422/

[13] 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.

[14] W. Li and J. Zhao, "TextRank Algorithm by Exploiting Wikipedia for Short Text Keywords Extraction," in *Proc. 2016 3rd International Conference on Information Science and Control Engineering*, 2016, pp. 683-686.

[15] A. Bougouin, F. Boudin, and B. Daille, "Topicrank: Graph-Based Topic Ranking for Keyphrase Extraction," in *Proc. The 6th International Joint Conference on Natural Language Processing*, 2013, pp. 543-551.

[16] A. M. Martinez and A. C. Kak, "PCA Versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228-233, Feb. 2001.

[17] S. Vogel, H. Ney, and C. Tillmann, "HMM-Based Word Alignment in Statistical Translation," in *Proc. The 16th International Conference on Computational Linguistics*, 1996, pp. 836-841.

[18] 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.

[19] 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.
[20] Y. Qian, C. Jia, and Y. Liu, "Bert-Based Text Keyword Extraction," *Journal of Physics: Conference Series*, vol. 1992, no. 4, p. 042077, Aug. 2021.

Wenting Hou is currently studying for the Master of Engineering (Engineering and Technology), Panyapiwat Institute of Management, Thailand. Her research interests are natural language processing, 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 Scienceand 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.