Journal of Ambient Intelligence and Humanized Computing (2021) 12:1023–1037
https://doi.org/10.1007/s12652-020-02119-3

ORIGINAL RESEARCH

Integrating machine learning and open data into social Chatbot for filtering information rumor

I-Ching Hsu[1] · **Chun-Cheng Chang**[1]

Received: 31 July 2019 / Accepted: 13 May 2020 / Published online: 20 May 2020

Abstract
Social networks have become a major platform for people to disseminate information, which can include negative rumors. In recent years, rumors on social networks has caused grave problems and considerable damages. We attempted to create a method to verify information from numerous social media messages. We propose a general architecture that integrates machine learning and open data with a Chatbot and is based cloud computing (MLODCCC), which can assist users in evaluating information authenticity on social platforms. The proposed MLODCCC architecture consists of six integrated modules: cloud computing, machine learning, data preparation, open data, chatbot, and intelligent social application modules. Food safety has garnered worldwide attention. Consequently, we used the proposed MLODCCC architecture to develop a Food Safety Information Platform (FSIP) that provides a friendly hyperlink and chatbot interface on Facebook to identify credible food safety information. The performance and accuracy of three binary classification algorithms, namely the decision tree, logistic regression, and support vector machine algorithms, operating in different cloud computing environments were compared. The binary classification accuracy was 0.769, which indicates that the proposed approach accurately classifies using the developed FSIP.

Keywords Machine learning · Chatbot · Cloud computing · Open data

1 Introduction

Social platforms are online websites for convenient and rapid information dissemination. They allow people to interact easily, which facilitates the spread of information. However, because anyone can post, information shared on social platforms, such as Facebook, Twitter, Instagram, and YouTube, is unverified. Such unverified information is the primary source of false or malicious rumors. Users often cannot judge the authenticity of information; this is primary problem of social platforms. Therefore, a service to provide credible information to verify this information may be advantageous.

Open data has become the mainstream of information technology in recent years. Open data is a kind of credible information provided by the government or well-known enterprises to promote the development of information systems. In this study, we used Open Government Data (OGD) as credible information to assist users in evaluating the authenticity of unverified information. Recently, Wuhan coronavirus (COVID-19) (Wikiquote 2020) has spread worldwide rapidly and caused panic. Some rumors about the Wuhan coronavirus circulating on Taiwan's social platforms have had a negative impact on epidemic prevention. For example, one such rumor is that using an alcohol bath can reduce fever and prevent the Wuhan coronavirus. This study aims to provide a credible OGD to interactively verify the information on social platforms.

Unverified information and open data continue to increase on the Internet. Machine learning and cloud computing can be used to analyze big data for intelligence and to improve the associated computing performance. In recent years, chatbots are emerging trends in the human–machine interface to provide interactive conversations with intelligence. This study proposed a general architecture that integrates Machine Learning and Open Data with Chatbot based on Cloud Computing (MLODCCC) to assist users evaluating the information authenticity in social platforms.

✉ I-Ching Hsu
hsuic@nfu.edu.tw

[1] Department of Computer Science and Information Engineering, National Formosa University, 64, Wenhua Rd., Huwei Township, Yunlin County 632, Taiwan

Food safety has always been a hot topic on the Internet. Facebook is a representative social platform where users can create fan groups and post articles. Numerous articles regarding food safety issues are spread via Facebook. Additionally, Facebook provides an Application Programming Interface (API) for Messenger that allows developers to easily create chatbots to facilitate interaction with users. We used Facebook to develop a food safety information platform (FSIP) to verify the feasibility of our proposed architecture that employs machine learning and open data with a chatbot and is based on cloud computing (MLODCCC). The primary purpose of the FSIP is to assist users in determining the authenticity of food safety related posts on Facebook.

The developed FSIP uses Spark cluster cloud computing ecosystem to promote computing efficiency and adopts machine learning associated with open data approach to enhance the credibility of food safety information in Facebook posts. There are two requirements must be met in the development of the FSIP. The first is the performance evaluation of machine learning combined with open data through Spark cluster cloud computing. It can present open data to combine with Facebook posts to form big data application. The second is the correctness assessment of the information provided by the open source combined with machine learning. It can facilitate to develop machine learning technologies and integrate into open data to promote intelligent capabilities for various Spark cluster cloud computing applications. This study makes three main contributions. First, the MLODCCC is proposed to integrate machine learning and open data with chatbot into Hadoop Spark cloud computing environment. Second, the MLODCCC is used to develop the FSIP that assists users to judge the authenticity of food safety related posts in Facebook. The FSIP carries out to integrate four emerging research areas: machine learning, open data, social chatbot, and cloud computing. Third, the performance evaluation of FSIP is analyzed and compared with various Hadoop Spark cluster cloud computing environment.

The remainder of paper is organized as follows. The next section presents some related studies. Section 3 presents an integrated Machine Learning and Open Data with Chatbot based on Cloud Computing (MLODCCC). Section 4 developed a Food Safety Information Platform (FSIP) based on the proposed MLODCCC. In Sect. 5, this study presents performance evaluation and experimental results. Finally, summary and concluding remarks are included.

2 Related work

Social networks play an important role in information sharing and dissemination. Users can instantly post information and read other users' posts on social platforms. The advantage of this fast information dissemination approach is that it can be easily and inexpensively obtained. In contrast, social platforms also promotes the rapid dissemination of fake information or rumor. Some studies (Gottifredi et al. 2018; Han et al. 2018; Shelke and Attar 2019; Zannettou et al. 2019) have focused on how fake information, malicious information, and rumors affect user behavior in social platforms. In Zannettou et al. (2019), authors presented an overview of the fake information ecosystem, including various categories of fake information, various different actors, and motivations. In Shelke and Attar (2019), authors analyze the rumor detection approaches and present the classification of current rumor detection approaches.

In recent years, rumor detection is an important research topic in social networks. The existing methods used to detect rumors can be divided into three approaches: networking, machine learning and deep learning (Sarah et al. 2020). Networking approach (Alrubaian et al. 2018; Kotteti et al. 2018) adopts social networking features, including the number of fans, posts, the reply content, timestamp, to evaluate information credibility. Machine learning approach (Habib et al. 2018; Li and Li 2019; Xuan and Xia 2019) uses statistical analysis techniques to automatically process content classification to detect rumors. In contrast, deep learning approach (Bhuvaneswari and Selvakumar 2019; Asghar et al. 2019; Srinivasan and Dhinesh Babu 2020; Xing and Guo 2019) focuses on employing neural network technologies for training and creating simulations to facilitate detection of rumors. In this study, we integrated machine learning approach and open data with social chatbot to assist users evaluating the information authenticity in social platforms. Machine learning technologies, binary classification, and Latent Semantic Indexing (LSI) are adopted to identify potential information and to determine similar information, respectively. Therefore, our study is the first one to integrate machine learning and OD technologies for rumor detection.

OGD are high-quality data provided by governments for free use by citizens. Compared with the unverified information on social networks, OGD are more credible. Many studies (Pereira et al. 2017; Wang et al. 2018; Zhao and Fan 2018) have explored how the quality of OGD can help developers reduce the time and cost of information system development. Machine learning technology is widely used in various information systems to enhance intelligence (Demarie and Sabia 2019; Kumar et al. 2019; Lee and Park 2019; Li et al. 2019; Liu et al. 2018; Park et al. 2020; Sharp et al. 2018; Xiao et al. 2018). Wireless sensor network (WSN) is the foundation of the Internet of Things (IoT). In (Kumar et al. 2019), authors survey various machine learning approaches applied to WSN. Industry 4.0 is a smart manufacturing that reduces manpower and costs. In Sharp et al. (2018), authors survey how existing research uses machine learning to achieve smart manufacturing.

With the advent of the big data era, unstructured data have been analyzed semantically in recent years. Zhu et al. (2016) proposed a new model through combining the algorithms of Word2vec and TF-IDF (Term Frequency–Inverse Document Frequency) for analyzing the unstructured data. Their study compared the accuracy of different classifications with an example of detecting damp-heat syndrome. With regard to the application of chatbot, UP et al. (Narendra et al. 2017) used a chatbot carrying with natural language to resolve the drawbacks of using keywords for searching in traditional knowledge management systems. Their ideas mainly focused on analyzing the strings entered by users to obtain the meanings to be expressed, with which to query the knowledge management system. The said drawbacks in traditional management system were resolved. In the field of food safety information related research, Geng et al. (2017) used web crawler tools based on the Text Density and Multi-factor Similarity Calculation to obtain the food safety related information from the Internet. The obtained information was further analyzed through text data analysis, allowing the government organizations to understand the recent hot events on food safety. The authorities concerned can strengthen their relevant food supervision.

3 Integrating machine learning and open data with Chatbot based on cloud computing

This study proposes an integrating Machine Learning and Open Data with Chatbot based on Cloud Computing (MLODCCC) architecture for supporting various missions involved in dealing with intelligent chatbot development. This study argues that MLODCCC can be adopted as a common scheme to integrate machine learning and open data uniformly using a fundamental chatbot. This architecture is depicted in Fig. 1, which comprises six modules, three online and three offline modules. The online modules, including Intelligent Social Application Module, Chatbot Module and Open Data Module, provide services to users at any time. By contrast, offline modules, including Data Preparation Module, Machine Learning Module and Cloud Computing Module, provide services used by the online modules only at specific times and cannot provide services to users. Because online and offline modules are involved when a request is sent from a local client to a remote server, the former must rely on the latter to process resource-intensive operations over the Internet.

In the training process, the machine learning module is used to design a predictive model that can be reused by the chatbot module once it is established. Therefore, the training process is not required every time the chatbot interacts with a user.

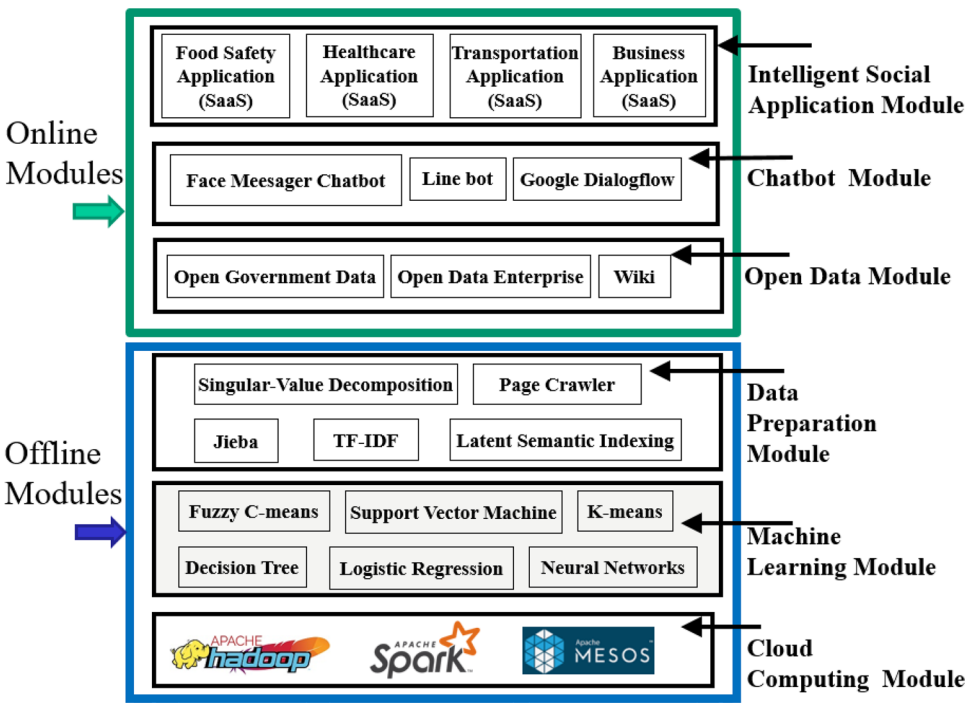

Fig. 1 The architecture of MLODCCC

3.1 Open data module

The "open" intent of open data is similar to some existing terminology, such as Open Source, Open Access, and Open Content. The purpose of open data is to promote the use, sharing, reuse and dissemination of data to facilitate developing information applications. Open data can be divided into three types: government, business, and individuals. The open data contained in this module is reliable information from government and business. The FSIP automatically accesses Taiwan government open data (Taiwan 2019) and imports to database.

3.2 Data preparation module

In the Data Preparation Module, collected data are preprocessed. The data source can be pure text data acquired from various sources. The collected data in this study included OGD, Facebook posts, and online news obtained using web crawlers.

Jieba (Jieba) is an open source software for automatic segmentation of Chinese word. There are four algorithms used in Jieba, namely DAG (Directed Acyclic Graph), Trie Tree, Dynamic Programming, and HMM Model (Hidden Markov Model). Term Frequency–Inverse Document Frequency (TF-IDF) is an algorithm used to calculate how important a word is to a document or Corpus. TF-IDF is a numerical statistic that is often used in text mining and data search (Yahav et al. 2019). Latent Semantic Indexing (LSI) is a method of information search (Phadnis and Gadge 2014). Generally, when searching for keywords would ignore the correlation among single words. However in practical applications, single words might actually one another be correlated. LSI can consider these correlations. Singular value decomposition (SVD) is used to decompose a matrix into three matrices (Onuki and Tanaka 2018). SVD is highly useful in machine learning to gather intelligence for various applications (Wang and Zhu 2017). The association between an article and a word is used in SVD to determine the category of an article.

3.3 Machine learning module

The Machine Learning Module utilizes numerous algorithms, including classification, grouping, and feature extraction algorithms. Using such algorithms, computer programs can learn the rules in data for use in decision-making and prediction. In this study, we used binary classification and feature extraction to identify potential food safety information and LSI to determine the similarities among food safety information, Facebook posts, and users' questions.

Binary classification divides a given data or elements into two categories according to the classification rules. The classification rules are usually based on the characteristics or attributes of the data for predictive classification. Binary classification belongs to a branch of machine learning study on statistical classification. It is a kind of supervised learning, which requires training data. Typical types of binary classification include Decision tree, Support vector machine, Neural Network, and so on.

3.4 Cloud computing module

Excessive data can lead to slow computing or even system failure. Cloud computing technology can solve these problems. In this study, the cloud computing technology used Spark for the computation of binary classification. Spark is an open source cluster computing platform designed for fast execution and high versatility (Yang et al. 2018). Spark boasts that its memory computing speed is 100 times that of Hadoop's MapReduce. Spark also supports multiple resource managers to help Spark perform the resource management for cluster computing, such as Standalone, YARN, and Mesos. In addition, Spark also provides a number of libraries, such as Spark SQL, MLlib, and GraphX.

3.5 Chatbot module

In the Chatbot Module, primarily keyword extraction and natural language processing are used to interact with users. Chatbots may be launched using various communication platforms. In this study, the chatbot provided by Facebook Messenger communication software was used. A chabot is a program that simulates the human dialogues for communicating with users. Chatbots are widely used on major instant messenger platforms, which provide easy-to-integrate webhooks facilitating chatbot development. At present, Chatbot has found its applications in various domains (Bates 2019; Okuda and Shoda 2019), for example health care, e-commerce, and finance. All of them have launched their chatbots to promote services.

3.6 Intelligent social application module

The member in the Intelligent Social Application Module is a specific domain social information system, which is a kind of Software as a Service (SaaS) (Hsu and Cheng 2015). The social SaaS is developed based on above modules, including Chatbot, Open Data, Data Preparation, Machine Learning, Cloud Computing Modules. The proposed MLODCCC provides a flexible infrastructure that social information system developer can dynamically add, replace, and remove components in each module. Each module contains multiple technologies, all of them providing a service suitable to the developer. Social information can be acquired from heterogeneous and distributed sources, including open data, social

platforms, and chatbots. The developed FSIP is one of the applications in the Intelligent Social Application Module.

4 Food safety information platform development

4.1 FSIP workflow

In this study, the Food Safety Information Platform (FSIP) was proposed to validate the feasibility of IMSCCCF. The open data module is used to design a predictive model that can be reused by the FSIP once food safety related open data is imported by manager. Therefore, the manager operation is not required every time the user send a new query. The following steps explain the message flow illustrated in Fig. 2.

Step 1 User.
1.1 With our FSIP, users can operate Facebook Messenger in a browser or on a mobile phone to obtain any food safety information they require.
Step 2 Manager.
2.1 The system manager imports the food safety related open data to the database.
Step 3 Messenger.
3.1 Messenger dialogues must be exchanged and delivered through the Messenger platform.
3.2 The data transmission and interaction between FSIP and Facebook Messenger is through the webhook server.
Step 4 Spark cloud computing.
4.1 Taiwan open government data is imported into the database.
4.2 The collected data are first pre-processed and then stored in the database.
4.3 The preprocessed data are subjected to binary classification using an algorithm to determine whether a Facebook post is related to food safety.
4.4 The binary classification algorithm is executed quickly through Spark's cloud computing platform.
4.5 Calculations are performed on the data received from the webhook server, and the results are displayed.

4.2 Data preparation

4.2.1 Data collection

The data sources can be various types of article file, which must be authorized and not subject to copyright or patent right restriction. In this study, we used six sets of food safety related OGD and news data from Taiwan. Taiwanese news covers a broad range of topics. We selected 47,900 health and safety articles, and 10,000 news articles selected randomly each of five topics: culture, politics, finance, sports and entertainment, and science and technology. Thus, we used 97,900 news articles for training data. At present, the FSIP is still in the experimental evaluation stage. In the future, OGD and news data related to food safety are planned to be updated quarterly.

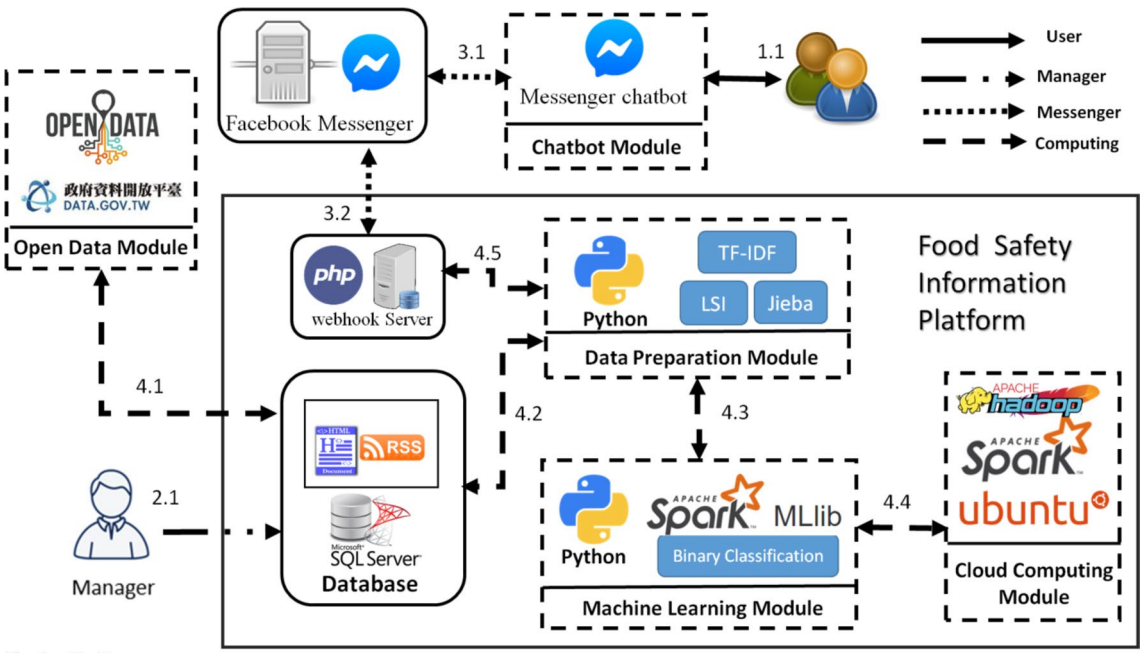

Fig. 2 The dataflow-oriented FSIP based on Spark cloud computing

4.2.2 Data preprocessing

Collected data must be processed before further use. The four steps in data preprocessing in this study were data cleaning, Chinese word segmentation, punctuation removal, and stop word removal. In data cleaning, HTML or XML files are parsed to extract clean data. In Chinese word segmentation, Chinese text was split into words, as shown in Fig. 3. Punctuation removal was required because punctuation symbols are unhelpful for extracting semantic meanings. Stop words are not necessary for text processing and were therefore removed.

4.2.3 TF-IDF

We used TF-IDF to calculate the weighted value of each word in each of the nearly 98,000 preprocessed articles. Each word–article pairing is known as a word–article. For example, assume calculations must be performed for 100 articles, and the TF-IDF value of a single word apple in an article d must be calculated. As expressed in (1), tf(t,d) represents the frequency at which a single word appears in an article. Assuming that the single word apple appears thrice in article d, the tf(t,d) of the single word apple is 0.33.

$$tf(t,d) = f_{t,d} \tag{1}$$

Next, we calculate idf(t,D). Assume that the single word apple appears in five of the 100 articles. According to (2), the idf(t,D) of apple is approximately 1.22184.

$$\text{idf(t, D)} = \log \frac{N}{1 + |\{d \in D : t \in d\}|} \tag{2}$$

The TF-IDF of apple in article d is calculated using tf(t,d) and idf(t,D). According to (3), the tfidf(t,d,D) = 0.3333 × 1.22184 = 0.407 for apple in article d.

$$\text{tfidf(t, d, D)} = tf(t,d) \times \text{idf(t, D)} \tag{3}$$

The TF-IDF of a single word–article can be calculated using the aforementioned steps. The TF-IDF for all other word–articles can be calculated by repeating these steps. The larger the TF-IDF of a word in an article is, the more relevant it is to that article.

4.2.4 Transforming articles to vectors with LSI

Machine learning requires a feature vector of data. Feature vectors usually consist of numerous numbers or other vectors. Articles must be transformed into vectors to be learned by our Machine Learning Module. In this study, the Latent Semantic Indexing (LSI) was used to transform the articles to vectors.

Through LSI, a word–article can be categorized into N topic categories according to certain requirements. LSI can obtain the final weighted values of N topic categories for each known word–article by using SVD on each word–article's feature vector. The workflow of transforming word articles to vectors with LSI is shown in Fig. 4. Each article was processed using LSI and SVD to convert it into a vector with 200 components. The nearly 98,000 articles were transformed into a 200 × 98,000 matrix, as shown in Fig. 5.

LSI transformed the articles into vectors, which were utilized for machine learning training and article similarity analysis. The following section introduces how an LSI model is created; detailed steps are presented in Fig. 6. In Sect. 4.4, we explain how the LSI model was used.

4.2.5 Creation of a LSI model

Figure 6 shows a flow chart of LSI model creation. First, open datasets must be preprocessed using word

Fig. 3 Chinese word segmentation via Jieba

```
seg_list = jieba.cut("九層塔含有黃樟素會導致肝癌？", cut_all=False)
print(" ".join(seg_list))
```

```
九層塔 含有 黃樟素 會 導致 肝癌 ？
```

```
seg_list = jieba.cut(" 網路流傳九層塔含有一種Eugenol成分(中文叫黃樟素)，會導致肝癌的說
print(" ".join(seg_list))
```

```
網路 流傳 九層塔 含有 一種 Eugenol 成分 ( 中文 叫 黃樟素 ) ， 會 導致 肝癌 的 說法
， 但 事實 為 九層塔 內 確實 含有 少量 Eugenol 成分 ， 不過 正確 中文翻譯 為 「 丁香
酚 」 ， 而 非 「 黃樟素 」 ， 其 英文名稱 為 Safrole 。 ( 2 ) 此外 ， 九層塔 中
雖然 含有 丁香酚 ， 但 其 毒性 低 ， 可以 快速 在 血液 及其 他 組織 中 代謝 ， 根據 2
006 年 世界衛生組織 食品 添加物 聯合 專家 委員會 公布 ， 丁香酚 一天 攝取 容許 量 ( A
DI ) 值為 2.5 毫克 / 公斤 ， 若以 60 公斤 重 的 成人 計算 ， 每人每天 的 安全 攝取量
為 150 毫克 。 因此 只有 在 大 劑量 攝入 丁香酚 時才 可能 對 生物體 產生 一些 健康 危
害 ， 目前 也 尚未 有 九層塔 致癌 的 研究 文獻 報告 ， 請 大家 勿 聽信 網路 謠言 ，
並 拒絕 轉傳 。 ？
```

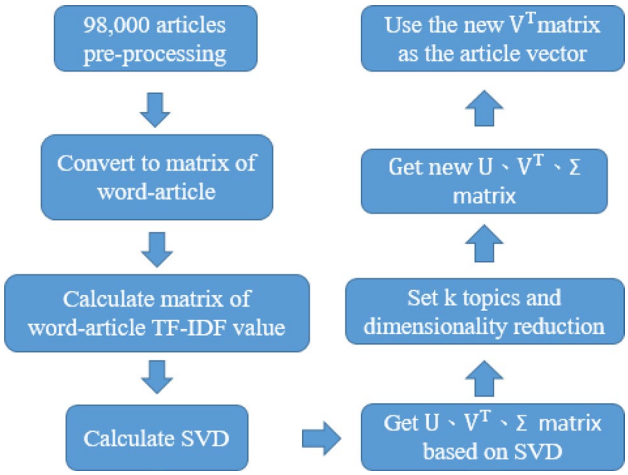

Fig. 4 Flow chart of transforming word articles to vectors with LSI

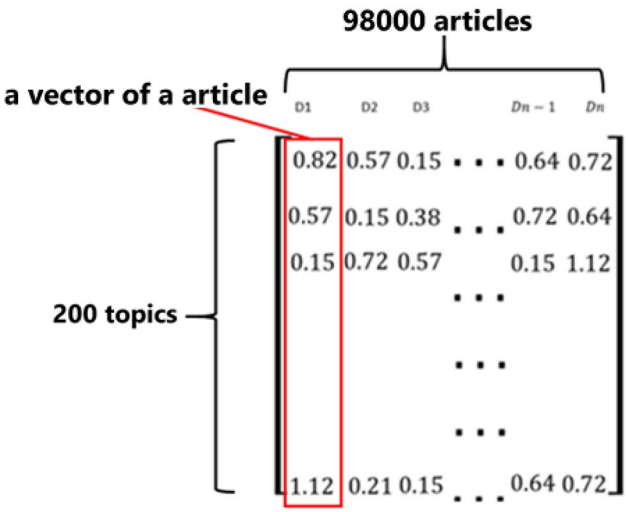

Fig. 5 Each article associated with topic is present in a matrix

segmentation, stop word removal, and synonym substitution. Next, open datasets are used to create a dictionary and calculate TF-IDF values. Finally, a topic is divided into 200 components, and the dictionary and TF-IDF values are input to the LSI algorithm to create the model.

4.3 Machine learning

This section describes the process of each stage of machine learning using binary classification algorithms. First, we introduce how data were divided into training, verification, and test data. Next, we specify how the aforementioned data were used to train and evaluate the binary classification model. Binary classification algorithms are diverse; we selected Decision Tree, Logistic Regression and Support Vector Machine (SVM) algorithms. The accuracy and performance of each algorithm were evaluated to determine which algorithm is most suitable.

4.3.1 Preparation of training data, verification data, and test data

After the transformation of word–articles into vectors, the vectors must be divided into two data fields, one for category features and another for numerical features. A category feature value of 1 indicated relevance to food safety, and 0 indicated no relevance. These features were subsequently regarded as a label and a feature, respectively, to generate data in LabeledPoint format, as shown in Fig. 7. The LabeledPoint data were further randomly divided into training, verification, and test data in a ratio of 8:1:1.

4.3.2 Stage of model training and evaluation

After division into the training, verification, and test data, three binary classification algorithms were used to evaluate and test the model, as shown in Fig. 8. We employed the area under the receiver operating characteristic curve (AUC) to evaluate the binary classification model (Atapattu et al. 2010). The possible binary classification parameter combinations were listed before model training began. After model training with one parameter combination was complete, the verification data were used to calculate the model's AUC; this AUC was added to an AUC list. These steps were repeated until all parameter combinations had been executed. Subsequently, the AUC list was sorted to identify the model with the highest AUC for use in the next stage. The test data were also used to calculate the model's AUC. If the AUCs calculated using test and verification data were the same, the model was considered not to be over trained and therefore the optimal binary classification model. The possibility of over training exists because the criterion used for selecting the test data is not the same as the verification data. If over training occurs during the testing phase, the model shall modify the parameter combination and re-execute the training phase.

4.4 System demonstration

The binary classification model and LSI model are combined to identify spurious food safety information in Facebook posts. The detailed process steps are shown in Fig. 9. First, the Facebook posts are processed and then transformed to vectors input to the binary classification model. If the output of the binary classification model is 1 (representing food safety relevance), then the vectors are input to the previously trained LSI model to obtain the weighted values of topic categories. The weighted values of topic categories are input to the LSI model, and their cosine

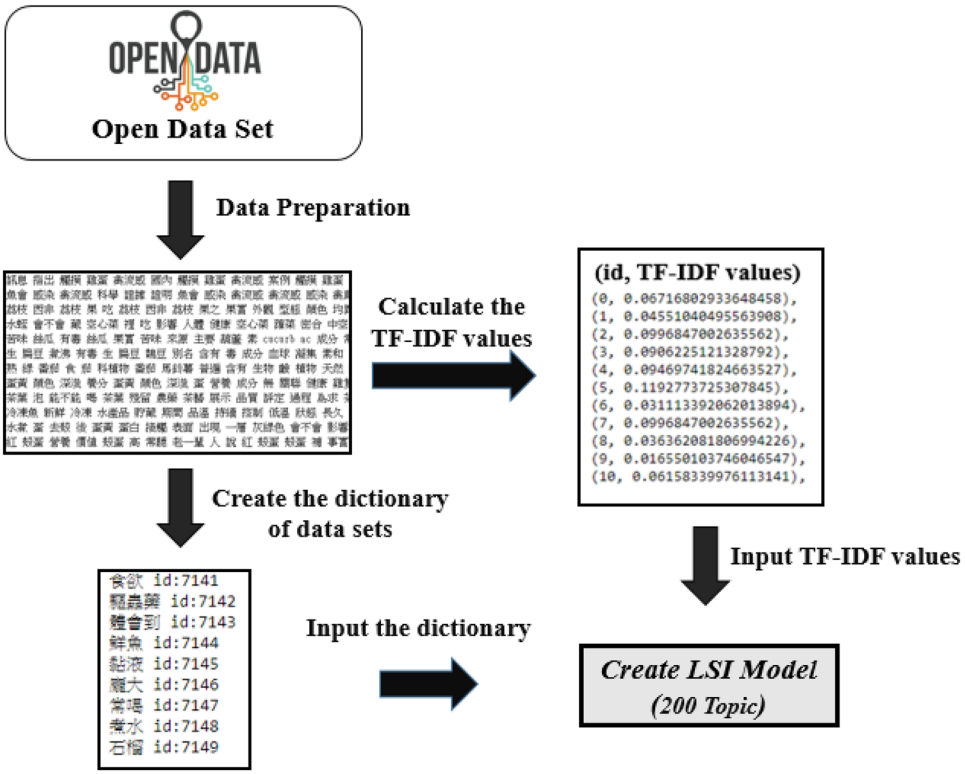

Fig. 6 Creation of LSI model

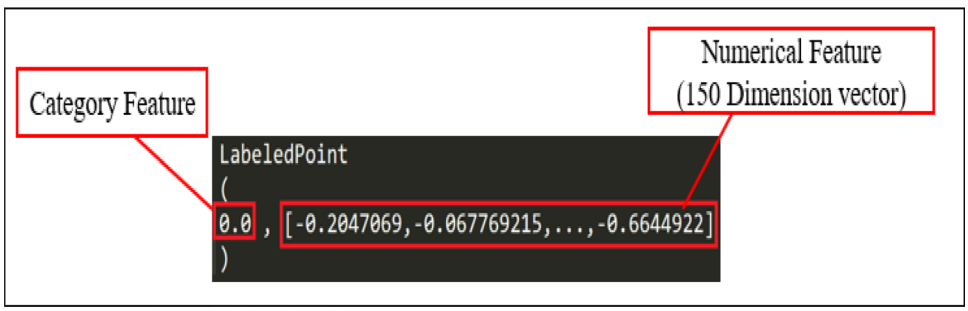

Fig. 7 LabeledPoint

similarity against the open datasets in the LSI model is calculated. The cosine similarity threshold was set to 0.8, based on the number of filtered articles, to obtain reliable comparison results. If the similarity is > 0.8 for an open data word–article, the open data and associated Facebook post are presented to the user. The user is reminded to view the post on Facebook.

Figure 10 shows the process of combining the LSI model with the chatbot. Most of the process is similar to that of combining the LSI model with the binary classification model. The main difference is that the platform for entering data is different. With a chatbot, users enter the questions; the received questions are transmitted to a messenger bot and then forwarded to a backend server for data preprocessing, LSI similarity analysis, and so on. Finally, based on the LSI model and cosine similarity operation, the four articles with the highest similarity are identified. When the similarity of the open data is > 0.8, the data are returned to the user through the Facebook Messenger bot.

This section explicitly demonstrates the applicability of the FSIP by describing how it can be used to provide credible open government data to users to judge the authenticity of unverified information in Facebook. The application of Facebook posts combined with open government data is shown in Fig. 11. When the user logs in to the FSIP using the Facebook account, the FSIP displays the latest Facebook post. A hyperlink (A) of open government data related to the content of the Facebook post (B) is displayed above the post. When the user has doubts about the content of Facebook post, they can click on the hyperlink to display the open data (C) provided by the government to help the user judge the authenticity of the post.

The interaction between Facebook messenger chatbot and user is shown in Fig. 12. The chatbot provides a menu

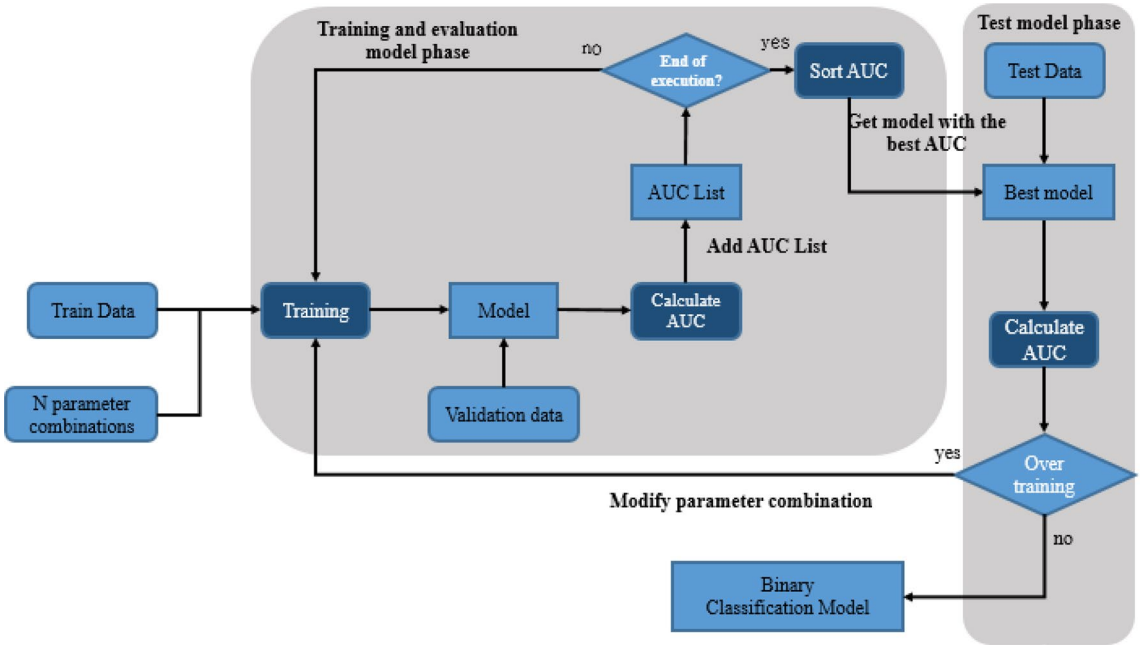

Fig. 8 Flow chart of model training and evaluation

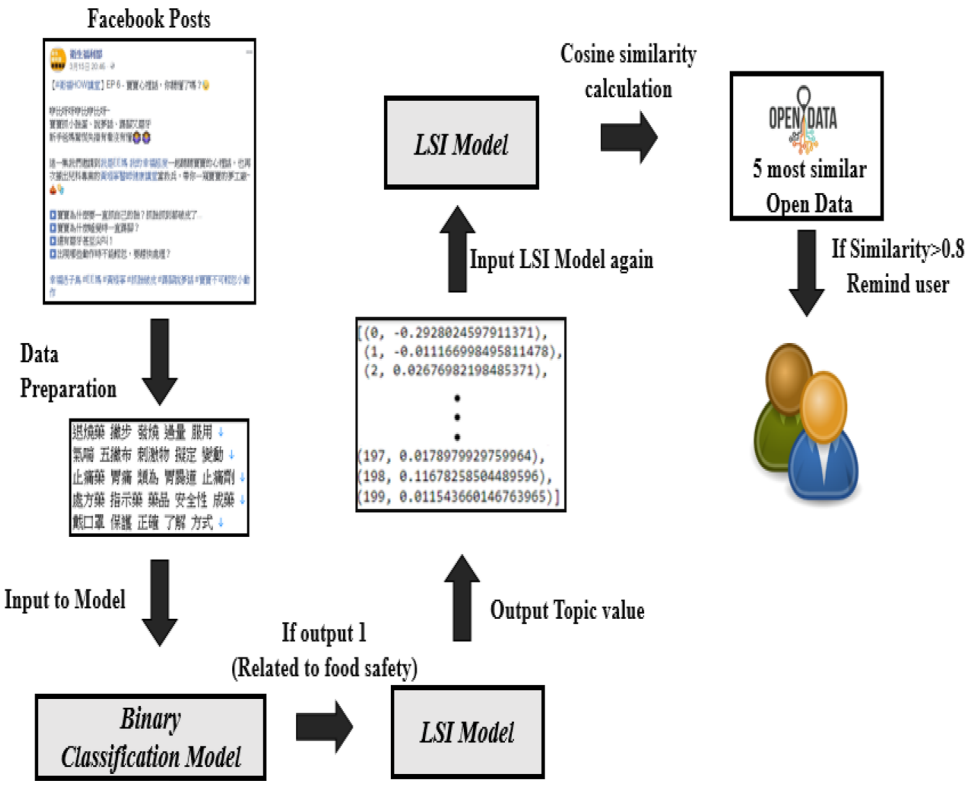

Fig. 9 Combination of LSI model and binary classification model

(A) to the user to select different interactive themes. The user clicks on the "Food Safety Information Search" button (B) to select food safety related topic to interact with the chatbot. The chatbot asks the user to select the category of food safety information (C). The subject selected by the user is food safety rumor (D). The chatbot asks the user to enter the food safety rumor (E). The user enters the food safety issue (F). The chatbot responds to the relevant open government data (G) to the user.

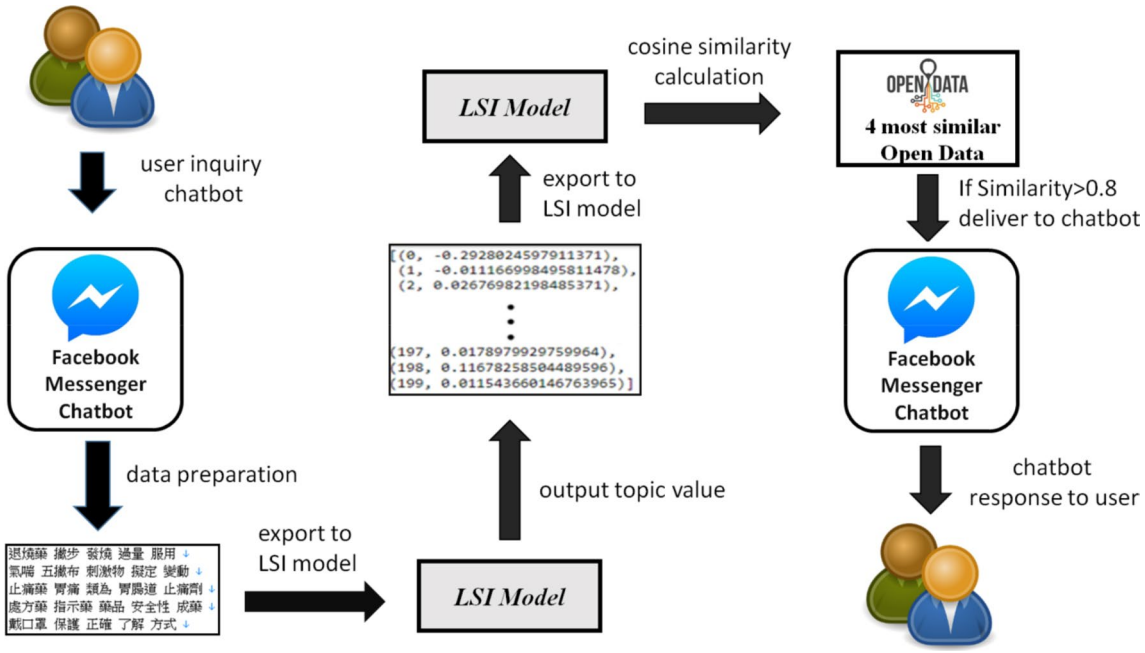

Fig. 10 Combination of LSI model and chatbot

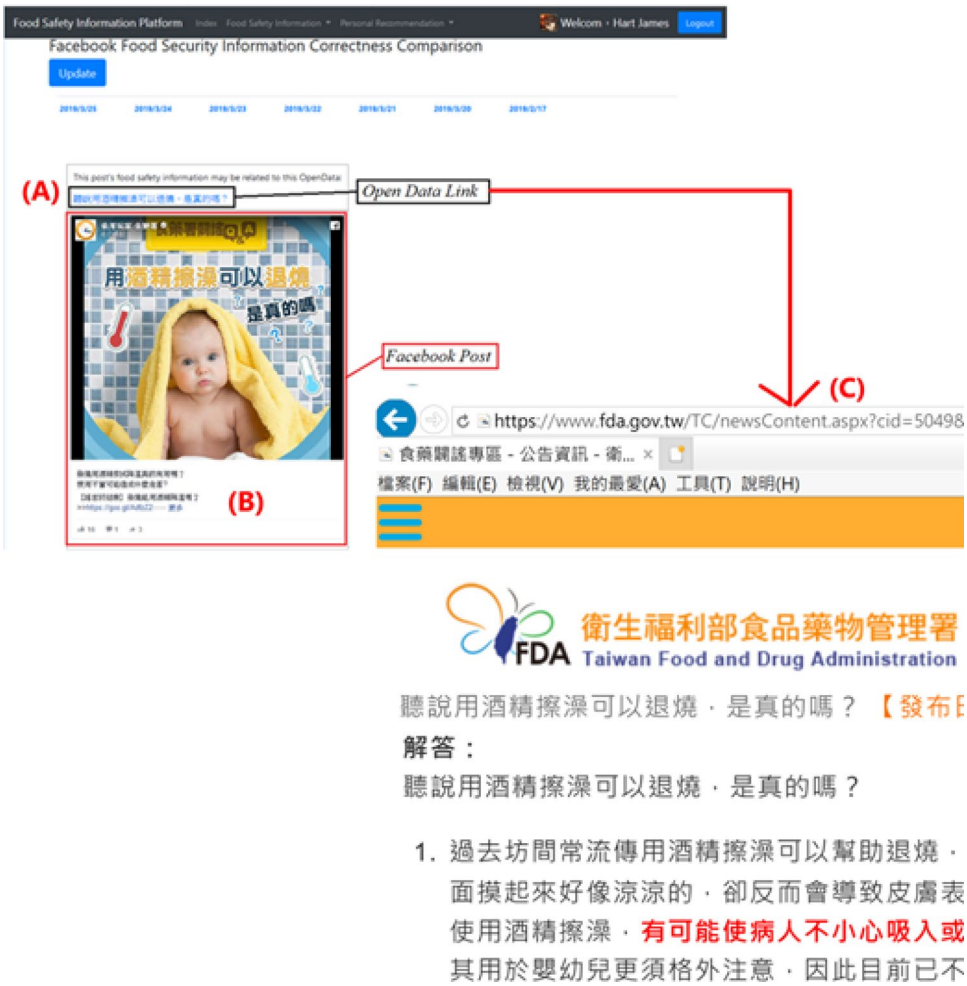

Fig. 11 Food safety correctness comparison

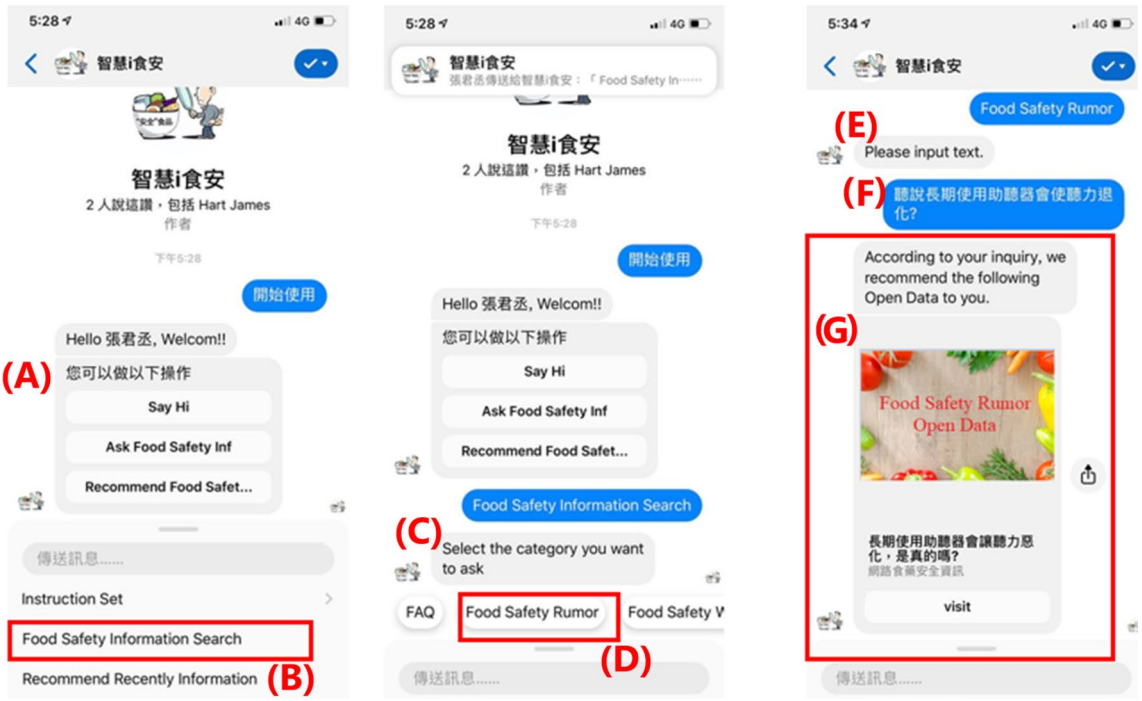

Fig. 12 Facebook messenger chatbot inquire open data through LSI

5 System experiment and results

Section 5.1 describes the Spark cluster environment used in the experiment, including Spark standalone mode and Spark YARN mode. Section 5.2 presents the performance evaluation of Decision Tree, Logistic Regression and SVM using Spark standalone mode and Spark YARN mode. Section 5.3 presents the probability prediction of Decision Tree, Logistic Regression and SVM using training data based on AUC. Section 5.4 presents the predicted results of a real case of Facebook fan posts.

5.1 Spark cluster environment

5.1.1 Computer specifications and cluster environment

In this study, we created a cluster environment with seven computers. Table 1 presents the hardware configuration of each computer, where one is used as the master node, and the other six, namely Data1 to Data6, are used as slave nodes.

In this study, Spark 2.2.0 and Hadoop 2.7.4 were built on seven servers respectively. One server was used as the mater, and the other six servers used as the slaves. Master server is responsible for assigning all the tasks to the six slave servers. All computational operations were dispatched by the master server to slave servers. Figure 13 shows Spark cluster environment.

Table 1 Computer hardware specification sheet

Host	CPU	Memory (GB)	HDD (TB)	OS
Master	i7-2600@3.4Gz	16	1	Ubuntu16.04LTS
Data1 (slave)	i7-2600@3.4Gz	16	1	Ubuntu16.04 LTS
Data2 (slave)	i7-2600@3.4Gz	16	1	Ubuntu16.04 LTS
Data3 (slave)	i7-2600@3.4Gz	16	1	Ubuntu16.04 LTS
Data4 (slave)	i7-2600@3.4Gz	16	1	Ubuntu16.04 LTS
Data5 (slave)	i7-2600@3.4Gz	16	1	Ubuntu16.04 LTS
Data6 (slave)	i7-2600@3.4Gz	16	1	Ubuntu16.04 LTS

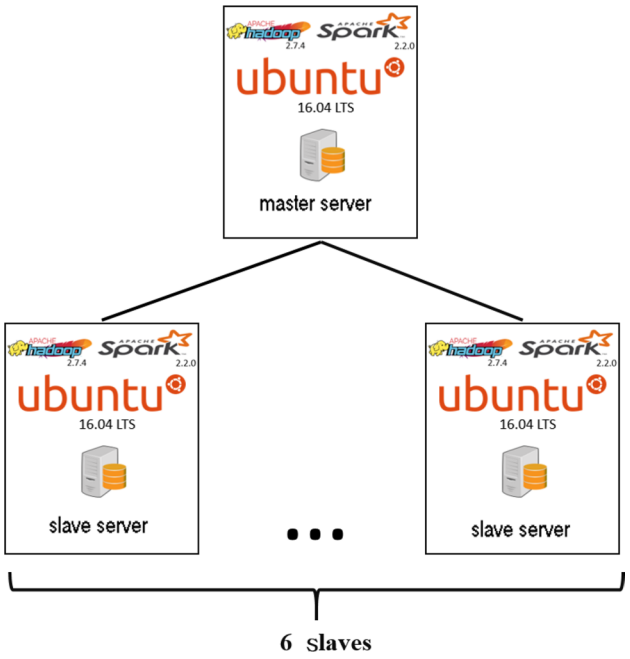

Fig. 13 Spark cluster environment

5.1.2 Spark standalone mode

The Spark cluster environment is composed of four components: Binary Classification, Spark Core, Cluster Manager and Cluster Manager, as shown in Fig. 14. Three binary classification algorithms used in Spark are provided in Spark MLlib, namely Support Vector Machine (SVM), Decision Tree and Logistic Regression. Spark Core is Resilient Distributed Datasets (RDD). Cluster Manager is Standalone in the Spark Standalone mode. File System is Hadoop Distributed File System (HDFS).

5.1.3 Spark YARN mode

Spark YARN mode is another mode in the Spark cluster environment. Its binary classification, Spark Core and file system are the same as the Spark Standalone mode. The difference between these two modes is the cluster manager. The Cluster Manager in Spark YARN mode uses YARN provided by Hadoop, as shown in Fig. 15.

5.2 Performance test

A performance test was used to compare the time spent by the Spark Standalone and Spark YARN cluster managers with the three binary classification algorithms for dataset sizes. 1 GB of data contains about 100,000 articles in our training dataset. The performance test results are shown in Fig. 16. The computational speed of the Decision Tree algorithm was higher than that of the other two algorithms, and for a small data volume, almost no difference was observed between the YARN and Standalone modes. For data volumes of $\geq$ 4.35 GB, the Standalone mode was clearly faster (by approximately 10–20 s) than YARN mode. For a small data volume, the computational speed of Logistic Regression in YARN mode was faster than that in Standalone mode (by approximately 5–10 s). Standalone mode was faster than YARN mode for data volumes $\geq$ 4.05 GB; at 5.85 GB of data, the speed was 12 s faster. The computational speed of the SVM was similar to that of Logistic Regression. YARN mode was 5–10 s faster than the Standalone mode for a small data volume; however, at $\geq$ 4.05 GB, Standalone mode was 5–10 s faster than YARN mode.

No considerable difference was observed in the computational performance of the algorithms between their use in YARN and Standalone modes. Standalone mode is likely to cause the loss of the child nodes for large datasets,

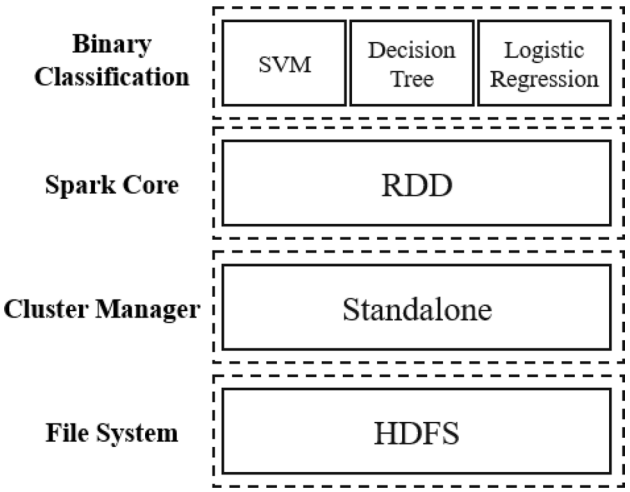

Fig. 14 Spark standalone mode

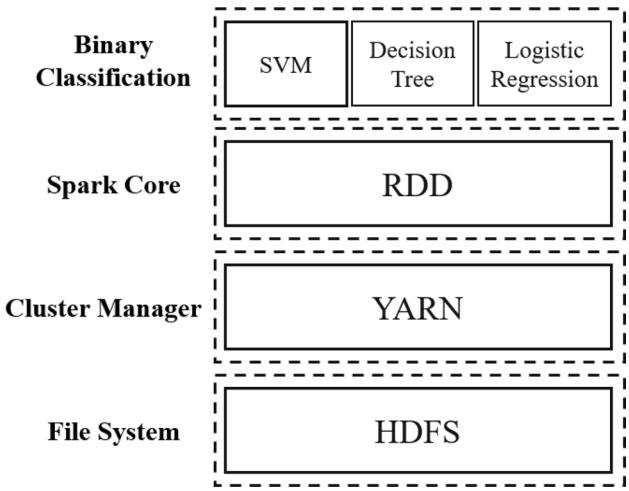

Fig. 15 Spark YARN mode

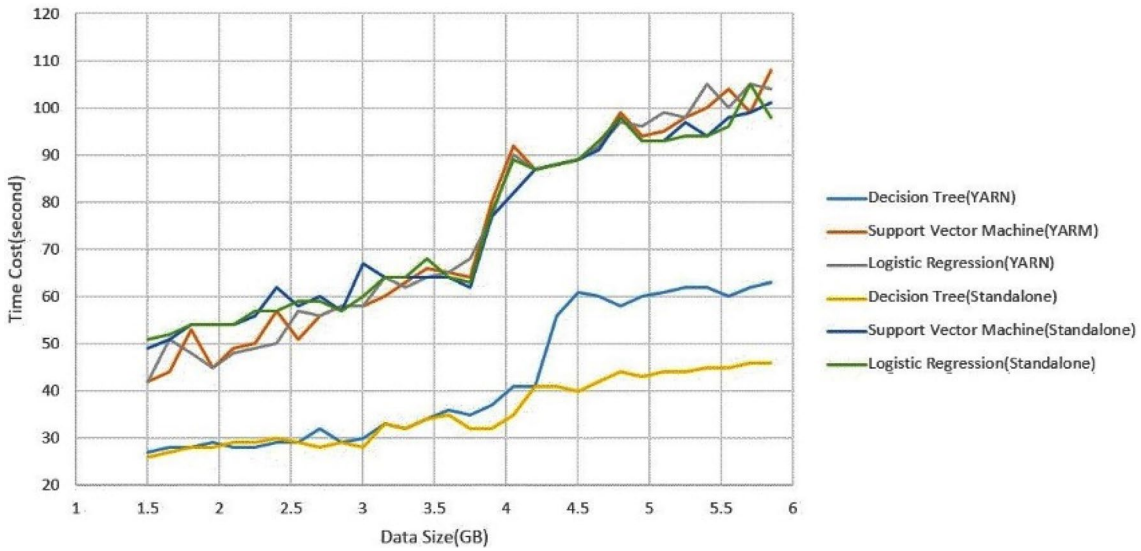

Fig. 16 The efficiency compared with different test model

Table 2 AUC test results of binary classification model

Train data	Decision tree	Logistic regression	SVM
1000 articles	0.8761	0.9012	0.9203
10,000 articles	0.9244	0.9601	0.9571
90,000 articles	0.9647	0.9662	0.9630

Table 3 Evaluation of the binary classification model in a real case

Method	True	False	AUC
Manually checked	145	116	N/A
SVM trained model	177	84	0.769

considerably reducing computational time. YARN mode is more stable than Standalone mode, with no problem of node loss.

5.3 Binary classification AUC test

Area under the Curve of ROC (AUC) was used to evaluate the binary classification models. AUC ranges from 0 to 1. A model with AUC = 1 has 100% predictive accuracy. When $0.5 < AUC < 1$, the results predicted by the model are worthy of reference. When $AUC \leq 0.5$, the predictive ability of the model is regarded as inadequate.

In this study, we trained the binary classification model using training data from 1000, 10,000, and 90,000 articles. The AUC test results of binary classification are shown in Table 2. First, 7900 articles were used as the test data for the model's AUC test. The trained model's AUC values calculated using three different binary classification algorithms were compared. When the training data size was 1000 articles, the performance of the decision tree was inferior to that of the other two algorithms, and the AUC of the model trained by the SVM reached 0.92. When the training data size was 10,000 articles, the AUC values of the model trained using the Decision Tree, Logistic Regression, and SVM were 0.924, 0.928, and 0.957. When the training data size was 90,000 articles, the AUC values of the model trained using the Decision Tree, Logistic Regression, and SVM were 0.964, 0.966, and 0.963. The AUC of all algorithms was greater for larger data volumes, but for small data sizes, the SVM was optimal. With larger data volumes, the Decision Tree and Logistic Regression also returned acceptable results.

5.4 A real case evaluation

This section evaluates a real case of Facebook fan posts based on the training test results in Sect. 5.3. Because of Facebook's security regulations, we could obtain only 270 fan posts for the study. The evaluation method was as follows. First, the FSIP called a Facebook API to obtain 261 fan posts from between March 3 and March 10, 2019. We manually examined whether these 261 posts contained food safety information. These results are presented in the second row of Table 3. The number of food safety–related posts was 145, and 116 posts were unrelated. The test results presented in Sect. 5.3 reveal that the SVM has the highest accuracy when the size of the training data is < 1000 articles; therefore, these 261 posts were classified by the SVM as containing or not containing food safety information. The results are shown in the third row of Table 3. The number of posts containing food safety information was 177, and 84 posts

were unrelated. In this real-world example, the AUC of the SVM was 0.769; therefore, the model is considered to have predictive value.

6 Conclusion

The rapid spread of rumors through social networks is a great concern. This study proposed a general architecture that integrates Machine Learning and Open Data with Chatbot based on Cloud Computing, called MLODCCC, to assist users evaluating the information authenticity in social networks. Food safety is crucial to daily life and is often discussed on social networks. Consequently, this study developed an FSIP for analyzing food safety rumors to verify the feasibility of its proposed MLODCCC architecture.

The FSIP is based on the Spark cloud computing environment and uses decision tree, logistic regression, and SVM algorithms for training and for generation of a binary classification model. Among the three algorithms, the prediction results of the SVM are superior when the number of training articles is small (<1000). When the number of training articles is large ($>90{,}000$), no considerable difference was observable among the algorithms. Additionally, a real case test of Facebook fan posts revealed that the binary classification accuracy was 0.769, which indicates that the proposed approach can effectively assist users in filtering information found in Facebook fan posts.

The MLODCCC proposed in this study is a modular and generalized structure that can promote the development and expansion of the state-of-the-art approaches. For example, the developed FSIP can be used not only in the field of food safety but also in other fields. Based on the MLODCCC architecture, we only need to expand the domain of training data. Another research direction is to use link open data (Khouri and Bellatreche 2018) to expand the application of open government data to improve the correctness of machine learning results.

References

Alrubaian M, Al-Qurishi M, Hassan MM, Alamri A (2018) A credibility analysis system for assessing information on twitter. IEEE Trans Depend Secure Comput 15(4):661–674

Asghar MZ, Habib A, Habib A, Khan A, Ali R, Khattak A (2019) Exploring deep neural networks for rumor detection. J Ambient Intell Human Comput. https://doi.org/10.1007/s12652-019-01527-4

Atapattu S, Tellambura C, Jiang H (2010) Analysis of area under the ROC curve of energy detection. IEEE Trans Wirel Commun 9(3):1216–1225

Bates M (2019) Health care Chatbots are here to help. IEEE Pulse 10(3):12–14

Bhuvaneswari ANG, Selvakumar S (2019) RumorDetect: detection of rumors in twitter using convolutional deep tweet learning approach. In: Paper presented at the 3rd international conference on computational vision and bio inspired computing, Coimbatore, India, pp 25–26

Demarie GV, Sabia D (2019) A machine learning approach for the automatic long-term structural health monitoring. Struct Health Monit 18(3):819–837

Geng Z, Shang D, Zhu Q, Wu Q, Han Y (2017) Research on improved focused crawler and its application in food safety public opinion analysis. In: Chinese Automation Congress, CAC 2017 Jinan, China. Institute of Electrical and Electronics Engineers Inc

Gottifredi S, Tamargo LH, García AJ, Simari GR (2018) Arguing about informant credibility in open multi-agent systems. Artif Intell 259:91–109

Habib A, Akbar S, Asghar MZ, Khattak AM, Ali R, Batool U (2018) Rumor detection in business reviews using supervised machine learning. In: Paper presented at the Proceedings—2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing, Kaohsiung, Taiwan

Han Q, Miao F, You L (2018) Rumor spreading model considering iterative spread on social networks. In: 18th IEEE international conference on communication technology, ICCT 2018, Chongqing, China. Institute of Electrical and Electronics Engineers Inc., pp 1363–1367

Hsu I-C, Cheng F-Q (2015) SAaaS: a cloud computing service model using semantic-based agent. Expert Syst 32(1):77–93

Khouri S, Bellatreche L (2018) LOD for data warehouses: managing the ecosystem co-evolution. Information 9(7):1–19

Kotteti M, Mouli C, Dong X, Qian L (2018) Multiple Time-Series Data Analysis For Rumor Detection On Social Media. In: IEEE International Conference On Big Data, Big Data 2018 Seattle, WA, USA. Institute of Electrical and Electronics Engineers Inc., pp 4413–4419

Kumar P, Amgoth T, Annavarapu CSR (2019) Machine learning algorithms for wireless sensor networks: a survey. Inf Fusion 49:1–25

Lee JH, Park SO (2019) Machine learning-based automatic reinforcing bar image analysis system in the internet of things. Multimedia Tools Appl 78(3):3171–3180

Li Z, Zhang K, Chen B, Dong Y, Zhang L (2019) Driver identification in intelligent vehicle systems using machine learning algorithms. IET Intel Transport Syst 13(1):40–47

Li S, Li Z (2019) Prediction of rumors wide-spreading on social media by logistic regression modeling: taking water resource pollution rumors spreading as an example. In: Paper presented at the 13th international conference on management science and engineering management, ICMSEM St. Catharines, ON, Canada

Liu S, Zhang L, Yan Z (2018) Predict pairwise trust based on machine learning in online social networks: a survey. IEEE Access 6:51297–51318

Narendra UP, Pradeep BS, Prabhakar M (2017) Externalization of tacit knowledge in a knowledge management system using chat bots In: 2017 3rd international conference on science in information technology: theory and application of IT for education, industry and society in big data era, ICSITech Bandung, Indonesia. Institute of Electrical and Electronics Engineers Inc., pp 613–617

Okuda T, Shoda S (2019) AI-based chatbot service for financial industry. Fujitsu Sci Techn J 54(2):4–8

Onuki M, Tanaka Y (2018) SVD for very large matrices: An approach with polar decomposition and polynomial approximation. In: 18th IEEE international conference on data mining workshops, ICDMW 2018 Singapore. IEEE Computer Society, pp 954–963

Park S-T, Li G, Hong J-C (2020) A study on smart factory-based ambient intelligence context-aware intrusion detection system using machine learning. J Ambient Intell Human Comput 11(4):1405–1412

Pereira GV, Macadar MA, Testa MG (2017) Delivering public value through open government data initiatives in a Smart City context. Inf Syst Front 19(2):213–229

Phadnis N, Gadge J (2014) Framework for document retrieval using latent semantic indexing. Int J Comput Appl 94(14):37–41

Sarah A, Alkhodair SHHD, Benjamin CMF, Junqiang L (2020) Detecting breaking news rumors of emerging topics in social media. Inf Process Manag 57(2):1–13

Sharp M, Ak R, Hedberg T (2018) A survey of the advancing use and development of machine learning in smart manufacturing. J Manuf Syst 48:170–179

Shelke S, Attar V (2019) Source detection of rumor in social network—a review. Online Soc Netw Media 9:30–42

Srinivasan S, Dhinesh Babu LD (2020) A neuro-fuzzy approach to detect rumors in online social networks. Int J Web Serv Res 17(1):64–82

Taiwan (2019) Taiwan open government data. https://data.gov.tw/ Accessed 5 May 2020

Wang D, Richards D, Chen C (2018) An analysis of interaction between users and open government data portals in data acquisition process. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11016 LNAI 184–200

Wang Y, Zhu L (2017) Research and implementation of SVD in machine learning. In: 16th IEEE/ACIS International conference on computer and information science, ICIS 2017 Wuhan, China. Institute of Electrical and Electronics Engineers Inc., pp 471–475

Wikiquote (2020) COVID-19. https://en.wikiquote.org/wiki/COVID-19. Accessed 5 May 2020

Xiao L, Wan X, Lu X, Zhang Y, Wu D (2018) IoT security techniques based on machine learning: how do IoT devices use AI to enhance security? IEEE Signal Process Mag 35(5):41–49

Xing F, Guo C (2019) Mining semantic information in rumor detection via a deep visual perception based recurrent neural networks. In: Paper presented at the 8th IEEE International Congress on Big Data, BigData Congress Milan, Italy

Xuan K, Xia R (2019) Rumor stance classification via machine learning with text, user and propagation features. In: Paper presented at the Proceedings—19th IEEE international conference on data mining workshops, Beijing, China

Yahav I, Shehory O, Schwartz D (2019) Comments mining with TF-IDF: the inherent bias and its removal. IEEE Trans Knowl Data Eng 31(3):437–450

Yang Y, Ning Z, Cai Y, Liang P, Liu H (2018) Research on parallelisation of collaborative filtering recommendation algorithm based on Spark. Int J Wirel Mobile Comput 14(4):312–319

Zannettou S, Sirivianos M, Blackburn J, Kourtellis N (2019) The web of false information: rumors, fake news, Hoaxes, Clickbait, and various other shenanigans. J Data Inf Qual 11(3):1–37

Zhao Y, Fan B (2018) Exploring open government data capacity of government agency: based on the resource-based theory. Govern Inf Quart 35(1):1–12

Zhu W, Zhang W, Li G-Z, He C, Zhang L (2016) A study of damp-heat syndrome classification using Word2vec and TF-IDF. In: Proceedings—2016 IEEE international conference on bioinformatics and biomedicine, Shenzhen, China, pp 1415–1420