Thai News Clustering Based on Machine Learning Approach

***Abstract*—There are many online news available on the web; however, not every news articles is telling a true story, some news websites are not trusted. Identification of trusted online news is a challenged task. Currently, there is no tool for checking the reliability of online news. In this research, we propose a methodology for online news trusted identification based on a text clustering technique. To find trusted news, collections of news articles from websites is needed as data mining sources. The collected news articles are then grouped according to the similarity of the words in the articles with the same meaning. The performance of the developed methodology is measured using sum squared error (SSE). The best performance is based on Hierarchical clustering algorithm with SSE lowest to 0.00.**

***Keywords*—machine learning, text mining, clustering, online news**

I. INTRODUCTION

Nowadays, the news on the Internet has grown rapidly and continuously that makes online news information huge. There are also social media allowing people to communicate with each other very quickly [1]. Social media become another channel for news broadcasts; it is easy to quickly share and spread the news both true and false news articles. News readers read the news without immediately knowing whether the news is a true or fake story, due to the lack of tools used to check the credibility of online news.

Text mining is a process for knowledge extraction from text documents. Since most text documents have no target labels, unsupervised or clustering methods are chosen as machine learning methods for text mining. Clustering can be used for trusted only news.

This paper proposes a methodology for Thai news clustering based on content similarity as a part of the future development trusted news process. K-means clustering and Hierarchical clustering are chosen for base clustering techniques. In addition, clustering performance between K-means and hierarchical clustering methods are compared.

The remainder of the paper is divided into the following sections: Section 2 gives details about literature review. Section 3 shows research methodology. Section 4 give results and discussion. Section 5 give conclusion remarks.

II. LITERATURE REVIEW

There have been some researchers studying on mining news online. Krishnamoorthy *et al.* [3] proposes a way to automatically categorize news articles from online news sources and label then based on their domains. His work uses the concept of natural language processing and machine learning. The clustering techniques used include rounded k-means clustering and incremental clustering. The clustering results of the proposed model is encouraging. These results can be used to compare the interests of different audiences in each news source. In addition, Nanayakkara and Ranathunga [4] proposed a technique for news aggregators helping readers to manage a large number of news by gathering them in one place with meaningful groupings. The news aggregator is available for English as well as other popular languages. Moreover, Shanker *et al.* [5] presented k-means integration with a hierarchical centroid shape descriptor for the division of brain tumors from multiple MR images.

III. METHODOLOGY

The research methodology for Thai news clustering in this paper consists the following steps: 1) data collection, 2) text extraction, 3) document indexing, 4) clustering, and 5) performance evaluation. The research methodology is shown as flowchart in Figure 1.

A. Data Collection

The data collection process is to prepare data as a source for future process. In this study, the data were received from three popular news agencies using DON-WebCrawler. There were 484 documents in the political section retrieved with five important features that were metadata, news title, news content, release date, and news link.

B. Text Extraction

In Thai word extraction, there are many tools available for word extraction. In this research, the tool used for word extraction was Thai Lexeme Tokenizer (LexTo) developed by the National Electronics and Computer Technology Center (NECTEC). Once the word segmentation has been completed, the Document Indexing metric is calculated.

C. Document Indexing

Document indexing is the process of converting natural language documents into forms that the computer can process. This will create a content representative document in the form of a Term Weighting Vector. In the case of document indexing, each document uses the Term Frequency-Inverse Document Frequency (TF-IDF) method as shown in (1) [5]

$$TF - IDF = TF \times IDF$$
$$TF_t = \frac{n_t}{N} \quad (1)$$
$$IDF_t = 1 + log\left(\frac{D}{d_t}\right)$$
$$TF - IDF_t = \frac{n_t}{N} \times \left[1 + log\left(\frac{D}{d_t}\right)\right]$$

where n_t is the number of words *t* that appears in the document. *N* is total number of words appearing in the document. *D* is total number of documents. d_t is number of documents with the word *t* appear.

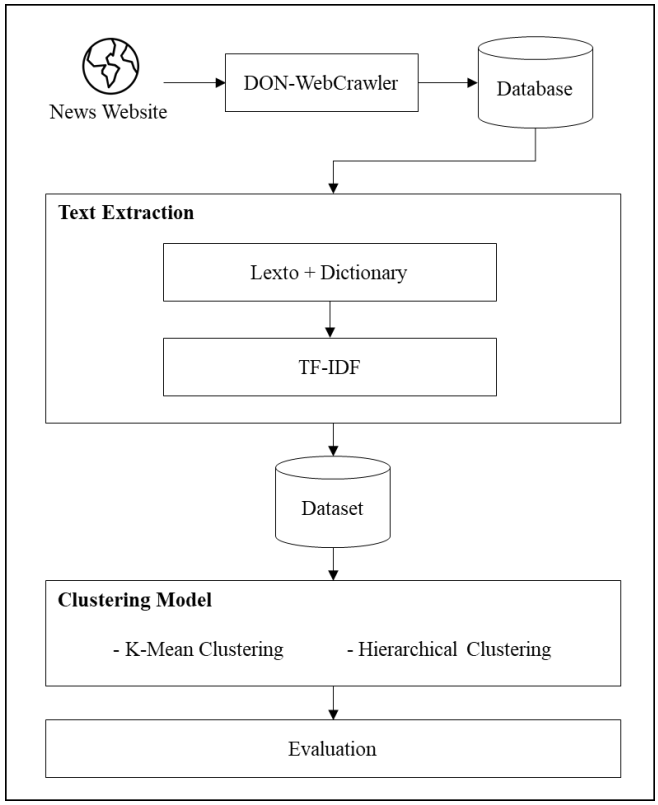

Fig. 1. Steps involved in Thai News Clustering

D. Clustering

There are many clustering techniques available to use. In this research, the two clustering techniques were used, which were K-Mean Clustering and Hierarchical Clustering.

E. Evaluation

For measuring the performance clustering results, this research uses the Sum of Square Error (SSE) metric as shown in (2). SSE finds the total error values between any data points in the cluster and the centroid of the cluster. This value represents the distribution of data. If SSE is high, the data within the cluster are very fragmented. On the other hand, if SSE is low, the data is well grouped together.

$$SSE = \sum_{i=1}^{k} \sum_{p \in c_i} d(p, m_i)^2 \quad (2)$$

where $d(p, m_i)$ is calculate distance between data functions. p is any data point in the c_i group and m_i is centroid of cluster.

IV. RESULTS

The clustering results are shown in Table 1. Comparing between k-means and hierarchical clustering, hierarchical clustering technique is outperformed k-means. However, both techniques have lowest SSE at five clusters. In addition, it is found that the best performance clustering technique is hierarchical clustering with SSE is 0.00.

TABLE I. THE TABLE SHOWS THE RESULTS OF CLUSTERING ALGORITHMS WITH THE K-MEAN AND HIERARCHICAL.

Cluster Technique	***Number of Group***	***SSE***
K-Mean Clustering	2	1
	3	0.9475
	4	0.7856
	5	0.3548
Hierarchical Clustering	2	0.9475
	3	0.5486
	4	0.3458
	5	0

V. CONCLUSIONS AND RECOMMENDATIONS

This paper presents a methodology for Thai news Clustering based on machine learning approach. First, the news data from websites are collected. Then, text extraction and document indexing are performed. Finally, clustering and evaluation processes are performed. It is found that for best clustering technique is hierarchical clustering, which is superior to k-means clustering. Further research is to use the text clustering technique for Trust online news.

REFERENCES

[1] C. Saini and V. Arora, "Information retrieval in web crawling: A survey," 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, 2016, pp. 2635-2643.

[2] A. Aun-a-nan and P. Meesad, “The Improvement of Web Crawler Perfomance for Online Thai News Content Collection and Extraction”, National Conference on Information Technology: NCIT, 2018, pp.121–127.

[3] A. Krishnamoorthy, A. K. Patil, N. Vasudevan and V. Pathari, "News Article Classification with Clustering using Semi-Supervised Learning," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, India, 2018, pp. 86-91.

[4] P. Nanayakkara and S. Ranathunga, "Clustering Sinhala News Articles Using Corpus-Based Similarity Measures," 2018 Moratuwa Engineering Research Conference (MERCon), Moratuwa, 2018, pp. 437-442.

[5] R. Shanker, R. Singh and M. Bhattacharya, "Segmentation of tumor and edema based on K-mean clustering and hierarchical centroid shape descriptor," 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO, 2017, pp. 1105-1109.

[6] J. R. Quinlan, “Induction of Decision Trees,” in Machine Learning, , 1986, pp. 81–106.