J. Inf. Technol. Appl. Manag. 28(5): 41~50, October 2021
https://doi.org/10.21219/jitam.2021.28.5.041

ISSN 1598-6284 (Print)
ISSN 2508-1209 (Online)

Misinformation Detection and Rectification Based on QA System and Text Similarity with COVID-19

Insup Lim* · Namjae Cho**

Abstract

As COVID-19 spread widely, and rapidly, the number of misinformation is also increasing, which WHO has referred to this phenomenon as "Infodemic". The purpose of this research is to develop detection and rectification of COVID-19 misinformation based on Open-domain QA system and text similarity. 9 testing conditions were used in this model. For open-domain QA system, 6 conditions were applied using three different types of dataset types, scientific, social media, and news, both datasets, and two different methods of choosing the answer, choosing the top answer generated from the QA system and voting from the top three answers generated from QA system. The other 3 conditions were the Closed-Domain QA system with different dataset types. The best results from the testing model were 76% using all datasets with voting from the top 3 answers outperforming by 16% from the closed-domain model.

Keywords : Misinformation Detection, Fake Information Detection, Qa System, Cosine Similarity, Machine Learning

Received : 2021. 02. 18. Final Acceptance : 2021. 10. 24.

* First Author, MSc Student, School of Business, Hanyang University, e-mail : insup88080@hanyang.ac.kr
** Corresponding Author, Professor, School of Business, Hanyang University, Professor, School of Business, Hanyang University, 222, Wangsimni-ro, Seongdong-gu, Seoul, 04763, Korea, Tel: +82-2-2220-1058, e-mail: njcho@hanyang.ac.kr

1. Introduction

At the end of 2019, WHO was informed of an unknown etiology detected in Wuhan City, Hubei province of China [WHO, 2020]. On 13 January 2020, the Ministry of Public Health, Thailand reported the first imported case of lab-confirmed novel coronavirus (2019-nCoV) from Wuhan [World Health Organization, 2020] and in an exponential rate with total reports of 57.8 million cases and over 1.3 million deaths on 22 November 2020 [World Health Organization, 2020].

The UN secretary-general launched the UN communications response initiative to combat the spread of misinformation in April 2020 [WHO, 2020]. At the World Health Assembly in May 2020, WHO Member States passed Resolution WHA73.1 on the COVID-19 response. The Resolution recognizes that managing the infodemic is a critical part of controlling the COVID-19 pandemic [WHO, 2020].

Under the COVID-19 infodemic, many fake news detection models have been developed, and the majority of the models were based on classification techniques. Classification techniques require tremendous amount of label datasets. Even there's a lot of true scientific information on the web, still it lacks false labeled information.

To resolve this problem, the proposed misinformation detection model is based on only using true information, Question Answering system (QA system) and text similarities.

2. Literature Review

Misinformation is defined as false information that is not created with the intention of hurting others [WHO, 2020a]. Even though it is not intended to hurt people, but to actually help them, still it is very dangerous for people to act on using unverified treatments against COVID-19.

2.1 Misinformation Detection and Rectification Systems

Classic misinformation detection models are based on supervised learning classification models, that need labeled datasets to identify whether they are true or not. Gilda [2017] tested TF-IDF with multiple machine learning algorithms, to identify fake news from Signal Media and OpenSource.co. Granik et al. [2017] used Naïve Bayes to classify fake news on Facebook news posts. Recently, after the COVID-19 outbreak, Elhadad et al. [2020] has built a machine learning model to classify fake COVID-19 information in Twitter by using voting systems of machine learning algorithms in multiple feature extraction conditions.

Rectification systems have been used based on QA systems, comparing the text similarity between the given answer and machine learning generated answer. Attia et al. [2018] have used an automatic short answer correction system based on course material to correct the wrong answer of the questions.

2.2 Question Answer Generation

Obtaining training data for QA system is time-consuming and resource-intensive, and current available QA datasets are limited to certain domains and languages [Lewis et al., 2019].

The question answering system is divided into two steps, that cloze generation and translation. Cloze generation is the process of masking noun or named entity, and cloze

translation transforms the masked sentence into question format [Lewis et al., 2019].

2.3 Open Domain Question Answering Model

There are two types of QA system, open-domain, and closed-domain. The Open-domain QA system uses all the contexts provided in the system to answer the question, finding the appropriate context with the semantic similarity between the question and each subset of contexts [Semnani et al., 2020]. On the other hand, the closed-domain QA system uses only one context for certain questions.

In this research both closed- and open-domain QA system, but the proposed model is more fitted into the open-domain system.

The QA system is fine-tuned upon the BERT-LARGE pretrained model, with 24 layers and 1024 hidden sizes, with 16 self-attention heads. The pretrained model is created by masking a word into a [MASK] token, and is trained to predict the masked word, allowing the model to understand the context [Devlin et al., 2018].

2.4 Word-Embedding

Word embedding is a technique of creating a word or sentence into a certain dimension sized vector. The tool used to vectorize the answers is BioBERT word-embedding module [Jangid, 2020], that is based on BioBERT [Lee et al., 2020], a model trained with 29 million PubMed articles with the same process of BERT. In this research the dimension of the word or sentence-embedding is 768, which is the same as the embedding size of BERT.

2.5 Text Similarity

Text similarity is used to compare how similar two text are to each other. Cosine similarity is one of the most widely used algorithm for text similarity. It is calculated between two vectorized words (or sentences) that is done from the word-embedding step [Li et al., 2013].

2.6 COVID-19 Datasets

Datasets were collected from Kaggle CORD-19 Competition [Wang et al., 2020], FAQs of WHO [WHO, 2020], CDC [CDC, 2020], CMU-MisCOV19 [Memon et al., 2020], CoAID [Cui et al., 2020], COVID-QA [Möller et al., 2020].

Datasets are divided into two types, scientific datasets and SNS, News datasets. Scientific datasets are mainly published journals or datasets from authorized organizations such as WHO and CDC.

CORD-19 dataset is a collection of published articles from major biomedical journals, mainly from PubMed Central, by well-known AI research communities to support the ongoing researches of COVID-19. It contains over 400,000 articles and 150,000 full texts about COVID-19, SARS-CoV-2, and related coronaviruses [Wang et al., 2020]. In this research, full paper is not used but the abstracts of the papers are used to build this model.

Frequently asked questions from WHO and CDC are also used as scientific datasets that are verified by these organizations. It provides basic information of COVID-19 and how to prevent it.

COVID-QA dataset is also in the category of scientific dataset, because the original contexts of Questions Answer data are published articles of COVID-19.

Next datasets are CMU-MisCOV19 datasets, and CoAID, CMU-MisCOV19 contains

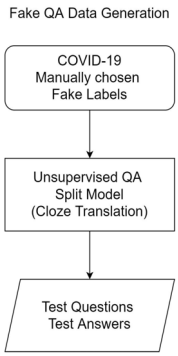
tweets datasets that is labeled with 16 different themes, such as true public health response, fake cure, conspiracy, irrelevant and so on [Memon et al., 2020].

CoAID is Covid-19 heAlthcare mIsinformation Dataset, that contains Tweets, retweets, social platform post (news) datasets [Cui et al., 2020].

3. Research Model and Implementation

3.1 Research Model

〈Figure 1〉 is the model of fake COVID-19 QA generation system, using the QA generation system based on Cloze translation. Datasets used in this model are CMU-M is COV19, CoAID dataset that is from tweets and news. By manually removing the contents duplicated, 153 False Question Answer sets were generated.



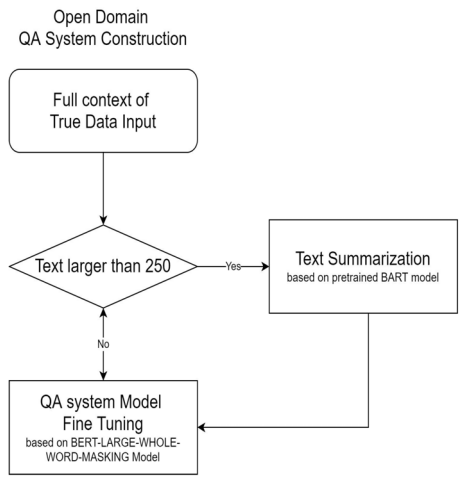
〈Figure 1〉 Fake Question Answer Generation

〈Figure 2〉 corresponds to the training system of the QA model. Since, data used for fine-tuning the QA model needs to be true, the full contexts are from CORD-19, combined FAQs of WHO and CDC, COVID-QA original contexts, true data of CMU-MisCOV19 and CoAID, directly mentioning the word in 〈Table 1〉 [Elhadad et al., 2020].

The scientific abstracts of CORD-19's word length were limited to 150 to 250 words, it is based on the word limit of PubMed Central (PMC), since most of the CORD-19 articles are sourced from PMC.

Text summarization was done for data that are over 250 words, such as the original text from COVID-QA dataset, to use as much various data as possible.

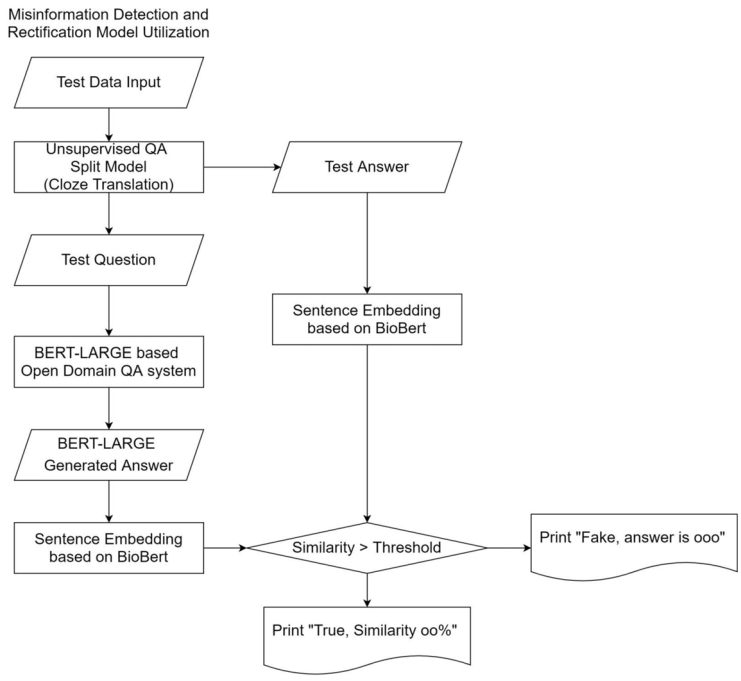
For the QA system, a lightweight wrapper for the deep learning library ktrain [Maiya, 2020], was used for loading the BERT-LARGE model, and build a QA model, that is based on an Open-Domain QA system.



〈Figure 2〉 Question Answering Model Fine-Tuning

〈Table 1〉 Keywords Equivalent to COVID-19

Keywords of COVID-19
• Coronavirus • Corona_virus • Corona-virus • Novel_Coronavirus • 2019-nCov • Novel-Coronavirus • NovelCoronavirus • 2019_nCov • COVID-19 • SARS-CoV-2 • Covid19



〈Figure 3〉 Misinformation Detection and Rectification Model Utilization

〈Figure 3〉 is the detection and rectification model utilization, first the data is put into the QA generation model, then the test question is put into the BERT-LARGE QA model that generates the answer. After the process, the test answer generated from the QA split model and the BERT-generated answer are converted into vectors by BioBERT-embedding library. These vectors are then compared with cosine similarity bringing out the value of 0 to 1, to determine whether the test answer is fake or not.

3.2 Model Implementation on Different Datasets

The model is implemented with 9 different conditions shown in 〈Table 2〉. From two different types of QA model and three different datatypes

The dataset types are divided into three categories. The scientific dataset, SNS and news dataset, and the aggregation of both datasets. The scientific datasets are the published articles and answers of the FAQs from WHO and CDC, the SNS and news datasets are mainly tweets and news that was contained in CMU-MisCOV, CoAID.

〈Table 2〉 Testing Conditions

Methodology	Dataset Types
Closed Domain QA system (base model)	Aggregated Data of Scientific, SNS and News
	Scientific Data
	SNS and News Data
Top BERT-LARGE answer with Open Domain QA system	Aggregated Data of Scientific, SNS and News
	Scientific Data
	SNS and News Data
Top Three BERT-LARGE answer with Open Domain QA system	Aggregated Data of Scientific, SNS and News
	Scientific Data
	SNS and News Data

There are certain questions that cannot be answered by the QA model. In this case, the QA system generated answer is labeled as "No Response". The evaluation of the model is done by excluding the "No Response" answers.

The Closed Domain QA system is the base model, that of finding the highest matching theme context and find the answer within the text.

Top BERT-LARGE answer is based on Open Domain QA system that generates the answers by searching the highest semantic similarity from the whole context.

Top Three BERT-LARGE answer QA system is similar to the explanation above, but finds the answer with top three highest semantic similarities, then determines the answer by choosing the majority by voting.

There are three rules for determining whether it is true or misinformation for the top three answer condition, after calculating the cosine similarity between each three answers and the test answer.

1. If the question has three or two answers with the same result, vote for the majority
2. If the question has two answers with a tie, pick the top answer, since it has the highest semantic similarity between the context and the question
3. If the question has only one answer pick the top answer as the result

For example, if answer 1 and test data has similarity of 0.9, answer 2 and test data has 0.8, answer 3 and test data has 0.4 and the threshold of determining the true or fake is 0.7, then the true is 2 and fake is 1, meaning the test answer is true.

4. Results

4.1 Accuracy and F1 Score

As shown in 〈Table 3〉 the closed-domain QA system had at most 0.592 accuracy with 0.362 F1-score, which was using the scientific datasets, and the lowest score with the SNS and News Datasets that had the accuracy of 0.394 and F1-score of 0.257. In this case, the similarity thresholds were very high, which are 0.9, compared to other testing conditions.

〈Table 3〉 Closed-Domain QA System Results

Dataset Types	Similarity Threshold	F1 score	Accuracy
Aggregated Data	〉0.9	0.351	0.564
Scientific Data	〉0.9	0.362	0.592
SNS and News Data	〉0.9	0.257	0.394

〈Table 4〉 shows the top Answer comparison between the BERT-LARGE QA system generation and the given test answer. There are questions that could not be answered, and the value with "Excluded" is the value with those unanswered questions removed. The highest score among all within the "Excluded" category, the accuracy was 0.724 from the scientific datasets with 0.793 similarity threshold, whereas the lowest was the SNS and News Dataset with 0.573 accuracy and 0.654 f1-score with a similarity threshold of 0.740.

Unlike the other two results table 5 indicates dataset with the highest accuracy is the aggregated data of both scientific, SNS and News data. It has an accuracy of 0.760 with F1 score of 0.758, the SNS and News data still got the lowest accuracy and F1 score of 0.741, 0.733 respectively.

〈Table 4〉 Top Answer BERT-LARGE QA System Results

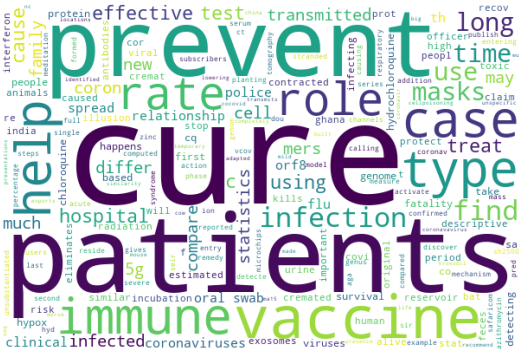
Dataset Types	Similarity Threshold	No Response	F1 score	Accuracy
Aggregated Data	〉 0.796	Included	0.715	0.722
		Excluded	0.719	0.717
Scientific Data	〉 0.793	Included	0.723	0.732
		Excluded	0.728	0.724
SNS and News Data	〉 0.740	Included	0.649	0.572
		Excluded	0.654	0.573

〈Table 5〉 Top Three Answer BERT-LARGE QA System Results

Dataset Types	Similarity Threshold	No Response	F1 score	Accuracy
Aggregated Dataset	〉 0.792	Included	0.753	0.764
		Excluded	0.758	0.760
Scientific Data	〉 0.792	Included	0.733	0.741
		Excluded	0.738	0.734
SNS and News Data	〉 0.791	Included	0.613	0.629
		Excluded	0.620	0.632

4.2 Analysis Of Correctly/Falsely Answered Questions

〈Figure 4〉 and 〈Table 6〉 show the most frequent words from correctly answered questions. According to 〈Figure 4〉 and 〈Table 6〉, cure, patients, vaccine, immune, and prevent were the most frequent questions that the QA system answered correctly

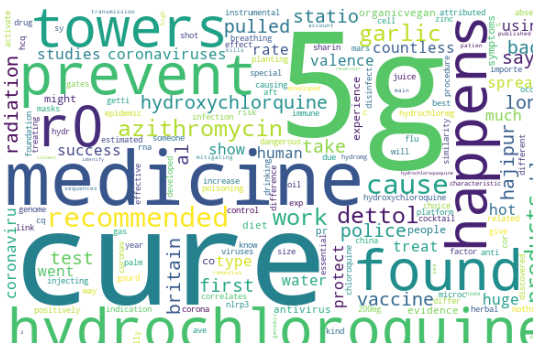


〈Figure 4〉 Wordcloud of Correctly Answered Questions

〈Table 6〉 Frequency of Word from Correctly Answered Questions

Most Frequent words	Word count
Cure	20
Patients	12
Vaccine	8
Immune	6
Prevent	6

〈Figure 5〉 and 〈Table 7〉 show the frequent words of falsely answered questions, which were 5g, cure, medicine, hydrochloroquine (correctly hydroxychloroquine), towers.



〈Figure 5〉 Wordcloud of Falsely Answered Questions

〈Table 7〉 Frequency of Word from Falsely Answered Questions

Most Frequent words	Word count
5g	10
Cure	7
Medicine	4
Hydrochloroquine	3
towers	3

From these results, the answers containing 5g, medicine were not answered correctly. For the word "cure", it appeared in both sides that 74% of answering the questions.

5. Conclusion and Limitation

The research aims at detecting and rectifying misinformation of COVID-19. Most of the

misinformation detection models are based on classification models, without correcting the false information, and the limitation of these models. First, it cannot indicate which part the wrong information is in the context, it may identify it as false, but it may not indicate which part is wrong. Secondly, there should be tremendous datasets that are labeled true or false to train the model. To solve these problems, the proposed model uses Question Answer system, Question Answer generator, text similarities and only true labeled datasets for training.

The goal of this study has been tested between 9 testing conditions. Comparison between closed-domain, and open-domain question answer systems, scientific dataset and SNS, News dataset, and choosing the highest semantic similarity or the highest three candidate answers for voting.

The top accuracy among all testing conditions was 76% that used aggregated dataset of scientific, SNS and news datasets for Open-domain question answering system with voting the top three candidate answers, outperforming the highest accuracy from closed-domain using scientific dataset by over 16%.

From the analysis of words the model has predicted the words including "patients", "vaccine", "immune", "prevent", on the other hand, questions including words like "5g", "medicine", "hydrochloroquine (hydroxychloroquine)", towers" were not predicted properly. For the word "cure", it was in both sides with almost the most frequent words in the questions. The question including "cure" was predicted properly 20 times, whereas it was wrongly answered for 7 times showing 74% of correct answering rate.

There are several limitations that can improve the result of this model. First of all, the scarcity of computing resource is one of the main problems. This led to using the text summarization to reduce the size of the dataset, even abstract is an abbreviated essential information of the article, still it is not as precise as it compared to the full paper. If there were enough computing resources, it would have been able to use the full published articles of COVID-19.

Secondly, unlike other fields of study, COVID-19 is a new subject that has only been around a year that lacks various dataset, such as formatted question and answer datasets that are verified to predict with high accuracy.

Thirdly, since this was a binary classification with one boundary of determining true or false, we do not know the degree of how much is true or how much is fake. If it was labeled in much precise standard, there could have been more possibility to improve the model.

Finally, creating a new word-embedding model specialized to COVID-19 would improve the model by having accurate vectors, that leads to more accurate text similarity calculation to determine true or false of the data.

References

[1] Attia, Z. E., Arafa, W., and Gheith, M., "An automatic short answer correction system based on the course material", *Int. J. Intell. Eng. Syst*, Vol. 11, No. 3, 2018, pp. 159-163.

[2] CDC, "Coronavirus Disease (COVID-19)", Centers for Disease Control and Prevention, 2020, https://www.cdc.gov/coronavirus/2019-ncov/faq.html.

[3] WHO, "Situation Report-1 21 January 2020", World Health, 2020, 251.

[4] Cui, L. and Lee, D., "CoAID: COVID-19 Healthcare Misinformation Dataset", arXiv preprint arXiv:2006.00885, 2020

[5] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K., "Bert: Pre-training of deep bidirectional transformers for language understanding", arXiv preprint arXiv:1810.04805, 2018.

[6] Elhadad, M. K., Li, K. F., and Gebali, F., "Detecting Misleading Information on COVID-19" *IEEE Access*, Vol. 8, 2020, pp. 165201-165215.

[7] Gilda, S., "Evaluating machine learning algorithms for fake news detection", In *2017 IEEE 15th Student Conference on Research and Development (SCOReD)*, 2017, pp. 110-115.

[8] Granik, M. and Mesyura, V., "Fake news detection using naive Bayes classifier", In *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, 2017, pp. 900-903.

[9] Jangid, J., "Overfitter/biobert_embedding", BioBERT-Embedding Github, 2020. https://github.com/Overfitter/biobert_embedding/blob/master/LICENSE.

[10] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining", *Bioinformatics*, Vol. 36, No. 4, 2020, pp. 1234-1240.

[11] Lewis, P., Denoyer, L., and Riedel, S.. "Unsupervised question answering by cloze translation, arXiv preprint arXiv:1906.04980, 2019.

[12] Li, B. and Han, L., "Distance weighted cosine similarity measure for text classification", *International Conference on Intelligent Data Engineering and Automated Learning*, Springer, Berlin, Heidelberg, 2013.

[13] Maiya, A. S., "Ktrain: A Low-Code Library for Augmented Machine Learning", arXiv preprint arXiv:2004.10703, 2020.

[14] Memon, S. A. and Carley, K. M., "Characterizing covid-19 misinformation communities using a novel twitter dataset", arXiv preprint arXiv:2008.00791, 2020

[15] Möller, T., Reina, A., Jayakumar, R., and Pietsch, M., "COVID-QA: A Question Answering Dataset for COVID-19", In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, 2020.

[16] Semnani, S. J. and Pandey, M., "Revisiting the Open-Domain Question Answering Pipeline, arXiv preprint arXiv:2009.00914, 2020.

[17] WHO, "Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation", WHO.Int,, 2020, https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation.

[18] WHO, "Let's flatten the infodemic curve. Who.Int, 2020a, https://www.who.int/news-room/spotlight/let-s-flatten-the-infodemic-curve.

[19] Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., ... and Mooney, P., "CORD-19: The Covid-19 Open Research Dataset, ArXiv., 2020.

[20] World Health Organization (WHO), "Frequently Asked Questions on novel coronavirus-update", 2020, https://www.who.int/csr/disease/coronavirusinfections/faq_dec12/en.

[21] World Health Organization, "COVID-19 weekly epidemiological update", 2020.

Author Profile

Insup Lim

Insup Lim has obtained his master's degree in Business Informatics from Hanyang University, and bachelor's degree in Industrial and Systems Engineering from Virginia Tech. His research interests include business analytics, machine learning, and natural language processing.

Namjae Cho

Dr. Namjae Cho is a professor of MIS at the School of Business of Hanyang University, Seoul, Korea. He received his doctoral degree in MIS from Boston University, U.S.A. He has published research papers in journals including Industrial Management and Data Systems, Computers and Industry, International Journal of Information Systems and Supply Chain, Journal of Data and Knowledge Engineering. He also published several books including "Supply Network Coordination in the Dynamic and Intelligent Environment (IGI Global)" and "Innovations in Organizational Coordination Using Smart Mobile Technology (2013, Springer)". He consulted government organizations and several multinational companies. His research interest includes technology planning and innovation, analysis of IT impacts, knowledge management, industrial ICT policy, design thinking, and the management of family business.