MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190269

Analysis of the Health Information Needs of Diabetics in China

Xiaofeng Zhou[a], Yuan Ni[a], Guotong Xie[a], Wei Zhu[a], Cai Chen[a], Tianhao Wang[b], Zhigang Pan[b]

[a] *Ping An Medical Technology, Shanghai, China*
[b] *Department of General Practice, Zhongshan Hospital affiliated to Shanghai Fudan University, Shanghai, China*

Abstract

According to the latest statistics of the China National Health Protection Commission, the prevalence of adult diabetics in China has reached 11.6%, and the number of patients has exceeded 114 million. Understanding the needs of diabetics and what kind of problems they are anxious about are crucial for doctors, hospitals and other health care providers, which can be used to ameliorate patient education services and help patients to improve their disease management skills. Hence we have conducted a study to analyze the questions about diabetes collected from a Chinese health website; the number of which is 151,589. We have divided these questions into 9 categories using a convolutional neural network. The shocking results showed that the questions about prevention only account for 1.23%. And a Chinese patented drug Xiaoke pill, the main component of which is glibenclamide, ranks fourth among the drugs the user cares most about due to the cheap price. However, patients know very little about its side effects.

Keywords:

Diabetes Mellitus, Machine Learning, Natural Language Processing

Introduction

According to the China Internet Network Information Center, the number of Internet users in China reached 802 million in 2018, and the Internet penetration rate was 57.7% [1]. The Internet has become an important carrier for consumers to express their health information needs and search for health information. Furthermore, patients with chronic diseases are more willing to share experiences and seek help via the Internet [2]. In order to meet the needs of the market, there are dozens of companies in China that started offering online disease counseling services for patients, by constructing a connection between patients and doctors through the internet. As a result, these sites have accumulated a large number of disease-related questions, from which we can mine valuable information about what the patients are concerned. Health providers like hospitals can make use of these questions raised by patients to improve their services such as patient health education, and patient follow-up to consumers.

In 2000, the International Diabetes Federation (IDF) predicted that the number of diabetics worldwide would reach 366 million by 2030 [3]. But just 10 years later, in 2011, diabetes data released by IDF showed that the number of people with diabetes worldwide had increased to 366 million [4]. The spread of diabetes is faster than expected. Moreover the number of patients with diabetes in China has ranked first in the world [5].

Nevertheless, people with diabetes usually have no medical background. It's natural for them to feel overwhelmed and scared when faced with diabetes, which is incurable. Hence studying their needs and analyzing their anxieties are fundamental to health care providers to comprehend their needs, which can break the information gap between the patients and health care providers. As a result, health care providers can provide better patient education services and improve self-management ability of diabetics. In order to achieve this goal, we have chosen to study diabetes-related questions on the internet to analyze the needs of diabetics.

In previous studies, Haihong Guo et al. analyzed hypertension-related questions, which studied a very limited amount of data of 2000 questions [6]; Zongcheng Jia et el. studied cancer-related questions with a data of 1000 questions [7]. As far as we are concerned, there is occasionality in the distribution of such a small amount of data, which in general cannot represent the reality. And with the development of artificial intelligence, some traditional machine learning algorithms like support vector machines (SVMs) and Naive Bayesian have been used to classify questions by feeding some features as Bag of Words (BOW) or TF-IDF extracted from the dataset into the algorithm [8], which is not using state of the art technology, meaning that better models can be used to improve the performance.

To analyze the diabetes-related questions, we used a Convolutional Neural Network to classify these questions, which is capable of capturing semantic level information. Based on the results of the classifier, we have further explored the hidden information in the data. This method can be generalized to analyze other chronic diseases by means of feeding labeled data to our model.

Methods

Data Collection

In order to reflect the real distribution of the diabetes-related questions, we designed a spider specific for a Chinese health website (http://www.39.net/) to crawl all the diabetes-related questions without any selection or filter, of which the total number is 151,589. With such a large amount of data, we believe that the results obtained are highly convincing.

Since these questions are all raised by patients, most of which don't have a medical background, these questions about diabetes are very colloquial and there are a lot of misspellings. For instance, Many patients may type "二甲双胍"(metformin) as "二甲双瓜"(metformin). And some consumers may merge several questions into one single question like "Do I have diabetes? Is it necessary to take medicine? Which treatment is better?". The diversity and complexity of the dataset is also a huge challenge for us, which put forward higher requirements for the robustness of our model.

Classification Schema

According to the classification schema of the references [6,9], combined with the specific characteristics of diabetes itself, all diabetes-related questions are classified into 9 categories: diagnosis, treatment, lifestyle, complication, maternity related, prognosis, health provider choosing, prevention, and others. The "others" category represents some description irrelevant to disease. Using such a schema, each question can be classified into one of the classifiers.

Word2vec

Word2vec is an efficient model for learning high-quality, distributed vector representations that capture a large number of precise syntactic and semantic word relationships [10]. Unlike traditional Bag of Words representations, words are projected from a sparse, 1-of-V encoding using one-hot encoding (here V is the vocabulary size) onto a lower and dense dimensional vector space via a hidden layer. In general, the dimensions of vectors are set between 100 and 500. In such dense representations, semantically close words are likewise close in Euclidean or Cosine Distance in the lower dimensional vector space. Since Word2vec uses the unsupervised learning method, it's very efficient and easy to use.

It has been shown that performance can be improved by training the Word2vec model using domain-specific data before training a downstream neural network [11]. To train our own word2vec model, we fed a large corpus of text into the model, including all the diabetes-related questions and a lot of other medical-related documents. In our model, the dimension of the vector is set at 300.

As a result, each word can be represented in a distributed vector, containing syntactic and semantic information. Therefore, we can calculated the similarity that can be represented by the cosine similarity between the 2 vectors correspondent. Here are some interesting results:

Table 1. Word Similarity

Word1	Word2	Similarity
糖尿病 (diabetes)	高血压 (hypertension)	0.62
糖尿病 (diabetes)	苹果 (apple)	0.18
糖尿病 (diabetes)	水果 (fruit)	0.17
糖尿病 (diabetes)	蔬菜 (vegetable)	0.12
水果 (fruit)	高血压 (hypertension)	0.14
水果 (fruit)	苹果 (apple)	0.63
水果 (fruit)	蔬菜 (vegetable)	0.87

As is shown in the table above, the semantic information is well captured by the model trained. Such a well-trained model can improve the performance of our neural network model.

Convolutional Neural Network

There are two, main deep, neural network architectures in NLP (natural language processing): convolutional neural network (CNN) and recurrent neural network (RNN). Generally speaking, CNN is mainly used for classification tasks like sentiment classification since sentiment is usually determined by some key phrases; RNNs is mainly used for a sequence modeling task like language modeling as it requires flexible modeling of context dependencies [12]. Since our task is to classify questions, CNN is chosen as our basic network architecture.

Convolutional neural network, originally invented for computer vision tasks, which uses convolving filters to extract local features from images, have also been shown to perform well in natural language processing. With the help of the Word2vec model, sentences can be embedded in a matrix. After that, sentences can be treated just as an image which is originally represented by a matrix.

In this study, we used a multi-channel CNN which is largely used in our other classification tasks, and have also been proven to achieve excellent results. The main model architecture is shown in Figure 1.

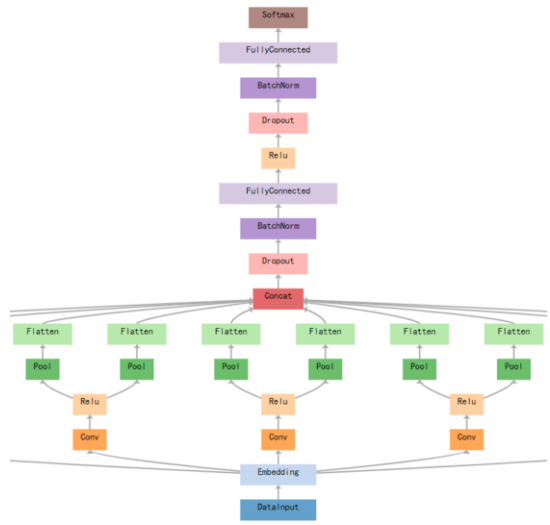

Figure 1. Main Model Architecture

Note: Each "Conv"(convolutional) module in this graph represents a convolving filter with different kernel size. Due to limitations of image size, not all the "Conv" module are presented in the figure. Max Pooling layer and Average Pool layer are used at the same time, represented by "Pool" module in the graph. To avoid overfitting problem, 0.5 is set as the dropout rate in the "Dropout" module.

Annotation and Training Procedure

To train a deep neural network, it's essential to feed labeled data into the deep model. We manually annotated 1000 questions, randomly selected from the dataset in the first place. Using such a small amount of data, we have been able to train a model that performs not very well but can already identify certain patterns according to the limited training data.

What we did next was to use the convolutional neural network to predict these unlabeled data and use the results as references to help us speed up the annotation process. The output of the model represents the probabilities that one sentence is classified into each category. Those sentences with low probabilities mean that the model fails to identify the patterns in the questions, which in most cases, represents that there is no similar pattern in the labeled data. So we need to annotate more data with new patterns for the model to learn. In this study, we defined the uncaptured pattern threshold at 0.7, and the captured pattern threshold at 0.95, meaning that if prediction probability of a question is below 0.7, we judge that the pattern of this sentence cannot be recognized by the model. In contrast, the prediction probability of a question above 0.95 means the sentence is well recognized.

Next, we manually annotated another 500 questions, randomly selected from the unrecognized questions. By repeating the training and annotation process above 4 times, we were able to train a well performing convolutional neural network.

Results

Neural Network Performance

We have annotated 3000 questions in total by means of the annotation and training strategy mentioned in the method section. All labeled data is divided into 2 sets, the training set occupying 80% for training our model, and the test set occupying 20% for evaluating our model, which is invisible to the model. The accuracy of our neural network has reached 96.7% in the test set, representing the performance of our convolutional model.

Using such an annotation and training strategy, we avoided annotating the same patterns all the time and were able to feed more new patterns to our neural network, which can largely improve the performance and robustness. Here is an example of the classification result to prove the robustness of our deep neural network:

- Question without misspelling:
  - 长期吃二甲双胍对健康都有些什么副作用？(What are the side effects of long-term consumption of metformin on health?)
- Question with misspelling:
  - 长期服用二甲双瓜有什么副作用(What are the side effects of taking metformin for a long time?)

Both questions are properly classified into the same class: treatment. The classifier works fine even if there is a spelling mistake in the sentence.

Classification Result

The classification result of the number of 151,189 questions is presented in Figure 2. We can conclude that the most concerned aspects of diabetics are diagnosis, lifestyle, and treatment. However, little attention has been paid to how to prevent diabetes, occupying just 1.23 percent, suggesting that patients rarely care about diabetes before actually getting diabetes.

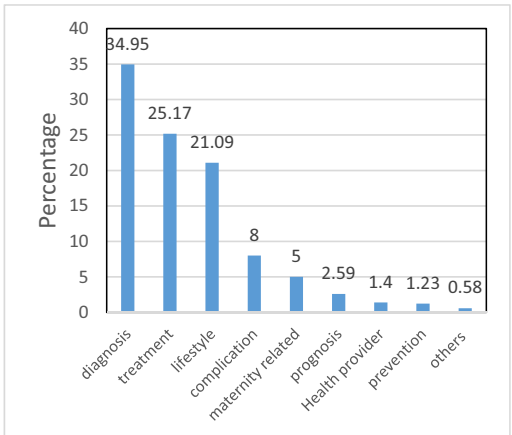

Figure 2. Classification Results

Study of Prevention Class

We carefully studied the questions of diabetes prevention. What we have found is that under most circumstances, people will start asking questions about how to prevent diabetes only when a relative or a friend has diabetes, here are some typical examples:

- 我的几个好朋友都得了糖尿病。怎么样才能预防糖尿病呢？(Several friends of mine have got diabetes. How can I prevent diabetes?)
- 身边有不少的亲戚朋友得了糖尿病，他们都很痛苦。所以我心里很担心，也很害怕。请问我该怎样预防糖尿病？(There are quite a few friends and relatives around me who suffer a lot from diabetes. So I am distressed and scared. How can I prevent diabetes?)

On the basis of the result, education about diabetes prevention needs to be spread to more people to help everyone recognize the importance of diabetes prevention.

Study of Lifestyle Class

Comparing questions before 2010 and after 2010 in Figure 3, the proportion of patients' questions about lifestyle increased from 13.71% to 22.59%, which shows a significant increase. A reasonable explication is that consumers are now more aware that lifestyle plays a very import role in diabetes.

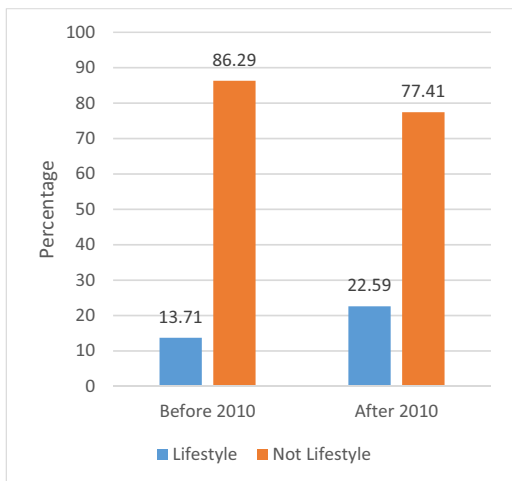

Figure 3. Comparison of Lifestyle Related Questions Before and After 2010

Knowing that diet plays a crucial role in diabetes, we conducted a further study about the most cited food by diabetics in the lifestyle class, shown in Figure 4. As for Chinese, eating has always been one of the most important things in daily life. Medical professionals can provide a better dietary recommendation for patients as a reference by means of the results obtained.

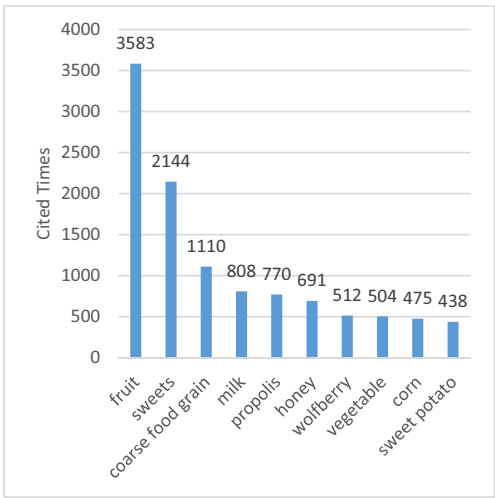

Figure 4. Top 10 Cited Foods

Study of Different Age Groups

It is also fundamental to analyze the different needs of patients of different ages, the result of which is presented in Figure 5. We can see significant differences in the aspects that people of different ages are concerned about.

Young people are more concerned about maternity-related issues, who are at an age suitable for childbearing. They may worry about problems like whether diabetes will be passed on to children or whether people with diabetes have fertility.

While older people are more concerned about complications and lifestyle to alleviate the impact of diabetes on life. Many people at this age are using oral medication or insulin therapy. They are more eager to know the drug usage, efficacy and adverse reactions.

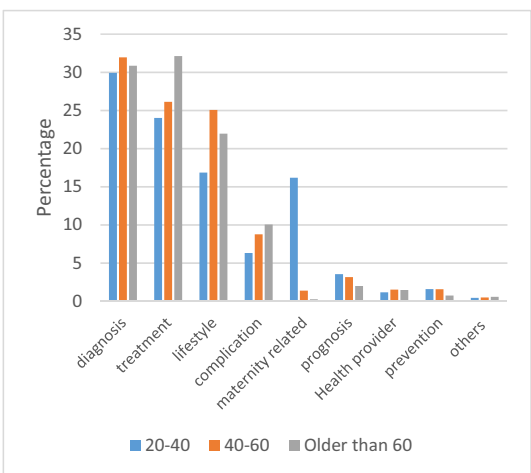

Figure 5. Classifcation Results of Different Age Groups

Study of Drugs

About the medication for diabetes, consumers raised more questions about insulin, metformin, gliclazide, and the Xiaoke Pill. Despite the import role of insulin in blood sugar control, the application of insulin is far from sufficient. Many diabetics, even medical staff, especially non-endocrinologists, have a variety of concerns and hesitation about the use of insulin, which suggests that diabetes professionals should teach patients and relevant medical staff more knowledge of insulin, improve their attitude towards diabetes.

The questioning rate of metformin is high, and that may be related to metformin being recommended as the first choice for diabetes treatment.

Xiaoke Pill is a kind of Chinese patent medicine, which is favored by many consumers because of the cheap price. Also, in China, many patients believe that Chinese traditional medicine has fewer side effects than western medicine, which may also be a reason why Xiaoke Pills are very popular in China. However, Xiaoke Pill contains glibenclamide, which can cause severe hypoglycemia and may lead to liver and kidney damage with a long-term application. Therefore, medical professionals should fully explain to consumers the side effects of Xiaoke Pills, and guide consumers to abandon the use of this drug.

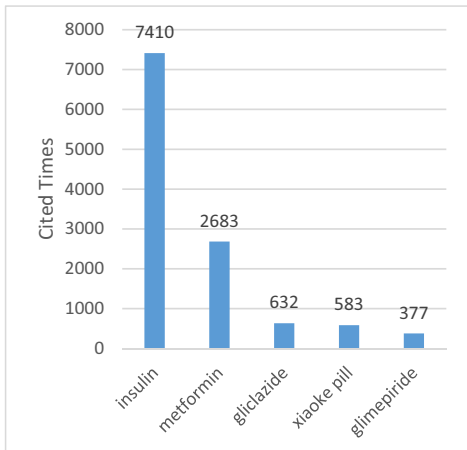

Figure 6. Top 5 Cited Drugs

Discussion

In this study, we have applied Word2vec and a state of art convolutional neural network to classify diabetes questions. In addition, we have used a more advanced annotating strategy, which can feed more patterns to the deep neural network to improve the performance. To achieve the same performance using the normal annotating strategy without pattern selection, we need a larger amount of annotated data.

Compared to the previous works presented in the introduction section, which tried to classify and analyze disease-related questions, the results of our work is more convincing. At the same time, we further explored more hidden information based on the results of the classifier.

Doctors, hospitals, and government medical institution can use the results of this analysis to better understand the needs of diabetics and discover the aspects in need of improvement.

For example, there are only a few problems related to the prevention of diabetes, meaning that people are not paying enough attention to the prevention of diabetes, which may be a reasonable cause of the growing number of diabetics in China.

There are also many shortcomings in our research. First, we have annotated 3,000 diabetes-related questions, which requires extensive work and takes a long time. Our model cannot be directly generalized to other diseases. To do a similar analysis of another disease, we have to redesign the

classification schema and annotate the new data. The classification schema we used is not perfect. For example, questions about sport, diet, and living habit are classified into the same class: lifestyle.

In the future, we should aim at improving the limitations of our work. Some more advanced deep neural networks like ELMO [13] and BERT [14] have achieved state of the art performance in many NLP tasks. Maybe these networks can be used to improve the annotating and training process. And, we can redesign a better classification schema that can be easily generalized to another chronic disease. We hope that more advanced natural language processing technologies can be applied to the medical domain.

Conclusions

In this study, we identified the diabetes information needs for consumers by analyzing a large amount of data. We have developed a classification schema for diabetes-related questions and used a convolutional neural network to classify the questions according to the schema designed. And, we found that the prevention of diabetes has not received enough attention in China and the popularity of Xiaoke Pill in China also needs to be reviewed.

Acknowledgements

This study is supported Department of General Practice and Department of Endocrinology, Zhongshan Hospital affiliated to Shanghai Fudan University.

References

[1] China Internet Network Information Center. China Internet Network Development State Statistic Report. 2018.8.
[2] Susannah F, Kristen P. Chronic Disease and the Internet. [http://www.pewinternet.org/2010/03/24/chronic-disease-and-the-internet/, Accessed November 25, 2018]
[3] International Diabetes Federation: Diabetes Atlas 2000. Brussels, International Diabetes Federation, 2000
[4] International Diabetes Federation: Diabetes Atlas 2011. Brussels, International Diabetes Federation, 2011
[5] International Diabetes Federation. IDF Diabetes Atlas, 8th edn. Brussels, Belgium: International Diabetes Federation, 2017.
[6] Haihong G, Xu N, Jiao L., Consumer health information needs and question classification: analysis of hypertension related questions asked by consumers on a Chinese health website, *Studies in Health Technology and Informatics* **216** (2015), 810-814.
[7] Zongcheng J, Yaoyun Z, Jun X, Xiaoling C, Yonghui W, and Hua X., Comparing Cancer Information Needs for Consumers in the US and China, *Studies in Health Technology and Informatics* **245** (2017), 126-130
[8] Haihong G, Xu N, Jiao L., Automatically Identifying Topics of Consumer Health Questions in Chinese, *Studies in Health Technology and Informatics* **245** (2017), 388-392.
[9] Zhang J, Zhao Y., A user term visualization analysis based on a social question and answer log, *Information Processing & Management* (2013) 1019–1048.
[10] Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS 2013.
[11] Yoon Kim. Convolutional Neural Networks for Sentence Classification. In Proceedings of EMNLP 2014
[12] Yin, Wenpeng; Kann, Katharina; Yu, Mo; Schütze, Hinrich. (2017). Comparative Study of CNN and RNN for Natural Language Processing. eprint arXiv:1702.01923.
[13] Peters, Matthew E.; Neumann, Mark; Iyyer, Mohit; Gardner, Matt; Clark, Christopher; Lee, Kenton; Zettlemoyer, Luke. (2018). Deep contextualized word representations. eprint arXiv:1802.05365.
[14] Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. eprint arXiv:1810.04805.

Address for correspondence

Corresponding author: Xiaofeng ZHOU, Ping An Medical Technology, 316rd Laoshan Road, Pudong District,Shanghai, China
Email: 9437852@qq.com.
Phone: +86-18672842160.