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

doi:10.3233/SHTI190350

Generating a Health Information Technology Event Database from FDA MAUDE Reports

Ethan Wang[a], Hong Kang[b], Yang Gong[b]

[a] *College of Natural Sciences, The University of Texas at Austin, Austin, TX, USA*
[b] *School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA*

Abstract

Patient safety events (PSEs), or medical errors, are major impediments to healthcare system safety. Health information technology (HIT) is expected to promote quality of care. Nonetheless, HIT also creates unintended consequences that concern patient safety consolidating a high-quality database of HIT events is essential to understanding their nature. Previous studies demonstrated the potential to use FDA Manufacturer and User Facility Device Experience (MAUDE) database to extract HIT events. In this study, we utilized classic and CNN models to extract HIT events from MAUDE. Both individual and combined models were evaluated on the test set, where the best model identified HIT events with ~90% accuracy and achieved a ~.87 f1 score. This model was capable of identifying HIT events in an HIT-exclusive database and serving as a quality and error check tool during event reporting. Moreover, the strategy of HIT event identification may scale in developing other PSE subtype-specific databases.

Keywords:

Patient Safety; Medical Errors; Information Storage and Retrieval

Introduction

Patient safety events (PSEs) are defined as "any event or action that leads to or has the potential to lead to a worsened patient outcome related to the event or action" [1]. Many PSEs can be attributed to medical errors - unintentional and preventable adverse effects due to poor care. Medical errors are the third leading cause of death in the United States, responsible for approximately 251,000 deaths annually [2]. In response, the Institute of Medicine published "*To Err is Human*," a paradigm-shifting report calling on states to provide public, standardized, and mandatory reporting systems [3]. Nonetheless, current practices of incident reporting are imperfect. While report collection has been heavily emphasized, little effort has been directed to analyzing reported events [4]. This is especially troubling because identification of common themes is crucial in identifying relevant issues to improve patient safety.

The adoption of health information technology (HIT) is critical in reducing medical errors [5]. The U.S. Department of Health and Human Services defines HIT as "the electronic systems health care professionals – and increasingly, patients – use to store, share, and analyze health information." HIT includes, but is not limited to, electronic health records (EHRs), personal health records (PHRs), and electronic prescribing systems [5]. HIT-enabled care yields improved data collection, data availability, cost efficiency, coordination of care, and risk analysis, thus reducing extraneous care and medical errors [6]. Hugely beneficial to healthcare, reliance on HIT is not without its faults. HIT unavailability, malfunction, and improper use may increase the likelihood of adverse events. Listed in the top 10 technology-hazards for healthcare by the Emergency Care Research Institute and comprising one sixth of all PSEs, reducing HIT errors will maximize patient safety [7, 8].

Collecting HIT events is the first step in identifying causes and preventing recurrence. The Agency of Healthcare Research and Quality (AHRQ) has spearheaded this task and has standardized PSE reporting, including HIT event reporting, in the Common Formats (CF). However, because the CF includes HIT as a set of contributing factors rather than as an event category, very few HIT event reports were archived [9]. As a result, most HIT events remain uncategorized or insufficiently categorized under only broader categories such as device or medical/surgical supply. In addition, many reporters leave fields blank due to the extensiveness and challenges in knowledge of patient safety reporting. These limitations make data collection on HIT events difficult. Therefore, an HIT-exclusive event database is in an urgent need.

The FDA MAUDE database is an extensive, online, and public resource with great potential for extracting HIT events. Updated once a month, the MAUDE searchable database contains a wealth of information, including manufacturer information, adverse event information, device information, and patient outcomes [10]. Although only an estimated 0.46~0.69% [11] of MAUDE reports are HIT-related, MAUDE contains over 6 million total reports from mandatory reporters (i.e. importers, manufacturers, and device user facilities) and voluntary reporters (i.e. patients and healthcare professionals) as of Aug 2018. This proportion produces challenges in the identification of HIT events [12]. The key problem is that standard classification algorithms are less effective on imbalanced datasets and that data pre-classification is necessary for suitable training [13]. This study expands on our previous study that generated a database containing 97% HIT events by using term frequency-inverse document frequency (TF-IDF), and biterm topic modeling (BTM) [14].

In this study, to improve on the low F1 score of our prior study, we upgraded the identification method by applying four classic classification algorithms (i.e., Logistic Regression, SVM, Bernoulli Naïve Bayes, and Random Forest), and a convolutional neural network (CNN). Both combined models and individual models were assessed to identify HIT events from MAUDE reports for database generation. An HIT specific database, the product of our model, will be helpful in identifying, preventing, and learning from HIT events. The identification strategy may also be scalable to other PSE subtypes.

Methods

Data pre-processing

A pre-filtered dataset of MAUDE reports was expert-reviewed to determine if the reports were HIT-related [11]. We then randomly categorized the reports to achieve a standard 70/10/20 training/validation/test set split.

Tokenization and word embedding

For the CNN model, Keras API with tensorflow backend was used to tokenize the unstructured text fields of each MAUDE report, and tokenized sequences were then padded to a maximum sequence length of 1,000 words [15]. For most deep learning text-classification models, word representation through feature vector classification is essential in maximizing accuracy [16]. Due to the absence of a large unsupervised training set to initialize word vectors, we used publicly available pre-trained word vectors provided by Global Vectors for Word Representation (GloVe). Based on word co-occurrence, GloVe word vectors are highly effective in characterizing word similarity. Our model used trainable GloVe vectors with a dimensionality of 50 [17].

For the classic models, the unstructured text fields of each MAUDE report were interpreted into a TF-IDF matrix using the scikit-learn python package [18]. TF-IDF weights word importance by calculating an inverse proportion of word frequency in a particular document and the percentage of all documents in which the word appears. TF-IDF is calculated through the following formula where $w_{i,j}$ is the TF-IDF weight, $tf_{i,j}$ is the number of occurrence of term i in document j, and N is the total number of documents.

$$w_{i,j} = tf_{i,j} * log(\frac{N}{df_i})$$

The words with higher TF-IDF weights are rarer, signifying greater importance in the document [19].

Classic model architecture

Following TF-IDF word embedding, four popular machine learning algorithms –SVM, Bernoulli naïve Bayes (BNB), random forest (RF), and logistic regression (LR) – were constructed using the sci-kit learn python package to classify the weighted text [18].

CNN model architecture

The CNN architecture was inspired by Kim's work [20]. We used Keras API with tensorflow backend to build our model [15]. After GloVe word embedding, we applied three 1D convolutional layers with kernel size 2, 3, and 4 respectively. Each 1D convolutional layer has 256 filters and was followed by a max-pooling layer to produce 256 features. To minimize overfitting, we applied 30% dropout to the max-pooled layer. The three outputs were then concatenated and flattened into a single layer that was fed to a dense layer of 1024 units with rectified linear units (relu) activation followed by 50% dropout. The output layer had one unit and was activated by a sigmoid function. Hyperparameters were tuned according to the model performance on the validation set.

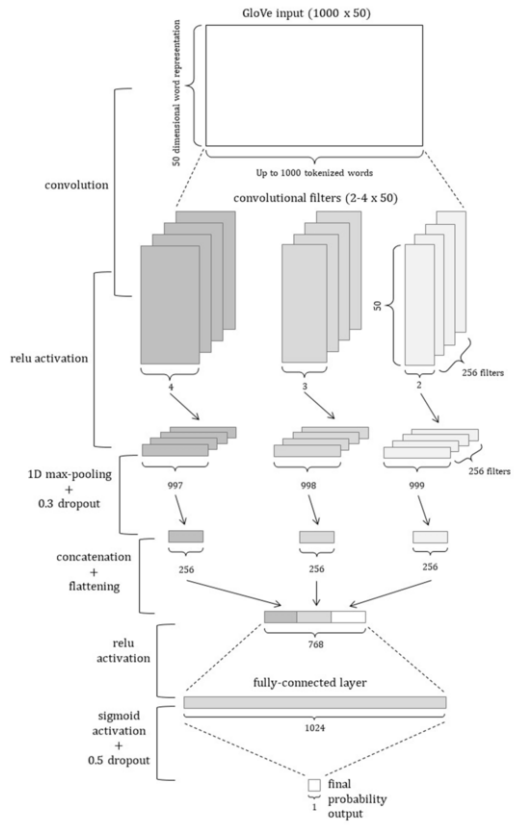

Figure 1: Illustration of CNN architecture for text classification. The figure depicts the three convolutional filter sizes (2, 3, and 4), max-pooling, concatenation, and final output layers used to classify a single document.

Combined model architecture

Among the five individual models, three with the best performance were applied to produce combined models. In each combined model, the probabilities of individual model were averaged to compute the probabilities of the combined model. If the averaged probability that the event is an HIT event equals or exceeds 50%, the combined model guesses 1 (HIT event), and if not, the model guesses 0 (non-HIT event).

Results

Dataset statistics

The dataset, containing 5,588 MAUDE reports, was generated by applying a keyword filter to recent nine-year MAUDE data (2008-2016) [11]. The dataset was relatively balanced between HIT - 2,192 (39.2%) and non-HIT – 3,396 (60.8%) PSEs. For model training, this balanced proportion is preferable to the 0.46~0.69% HIT event proportion in the MAUDE database as a whole [11]. In addition, we extracted 18,132 individual tokens from the 5,588 total reports.

CNN hyperparameter tuning

We tuned numerous CNN hyperparameters including batch size, learning rate, dropout, kernel size, max-pooling pool size, embedding dimensions, maximum token number, and

maximum sequence length. Increasing maximum token number and sequence length improved model accuracy. Thus, a maximum sequence length of 1,000 and maximum token number of 20,000 were used to incorporate as much training data as possible. Multiple dropout layers of 30-50% performed optimally, and max-pooling the product of each convolutional filter to an output of length 1 outperformed smaller max-pooling pool sizes. We found that batch size and learning rate had a negligible effect on model performance. Increasing embedding dimensions beyond 50 failed to improve model performance while increasing training time as well.

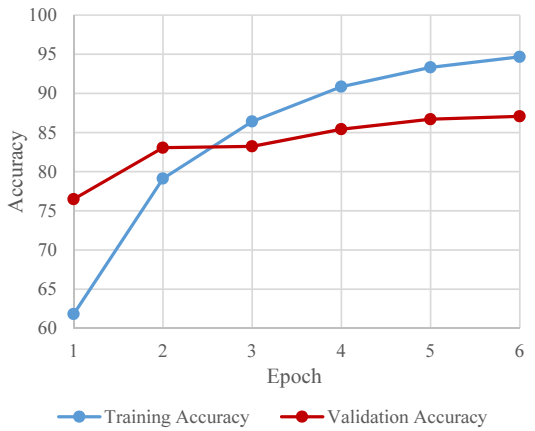

Figure 2: Training and validation accuracy of the CNN model over 6 epochs

While training the CNN model, we found that the rate at which validation accuracy improves decreases significantly around the 2th epoch, and fails to improve consistently past the 6th epoch (Figure 2). Moreover, training accuracy continues to improve past the 6th epoch, signifying overfitting of the training data. We found that the CNN model performed best on the test set when trained for 6 epochs.

Individual and combined model performance

Table 1: Accuracy, Precision, Recall, and F1-score of each individual and combined model on test set

Methods	Accuracy	Precision	Recall	F1
SVM+LR+CNN	.9012	.8796	.8606	.8700
SVM+LR	.8965	.8838	.8413	.8621
SVM+CNN	.8901	.8595	.8534	.8564
LR	.8895	.8828	.8366	.8591
LR+CNN	.8864	.8582	.8438	.8509
SVM	.8860	.8584	.8584	.8584
CNN	.8836	.8463	.8522	.8493
BNB	.8529	.8033	.8440	.8232
RF	.8467	.8690	.7186	.7885

*SVM = support vector machine, LR = logistic regression, CNN = convolutional neural network, BNB = Bernoulli naïve Bayes, RF = random forest

As shown in Table 1, combined models generally outperformed individual models by all metrics. LR+CNN was an exception to this trend. LR, the best performing individual model, outperformed LR+CNN. Among the other three combined models, SVM+LR+CNN outperformed SVM+CNN by all metrics and outperformed SVM+LR on all metrics except precision. Differences in the classification algorithms behind individual models may explain why the combined models performed better. One model may misclassify reports accurately classified by another model. Thus, averaging probability predictions can help mitigate the erroneous predictions of individual models by factoring in other models as well.

Effect of data size on model performance

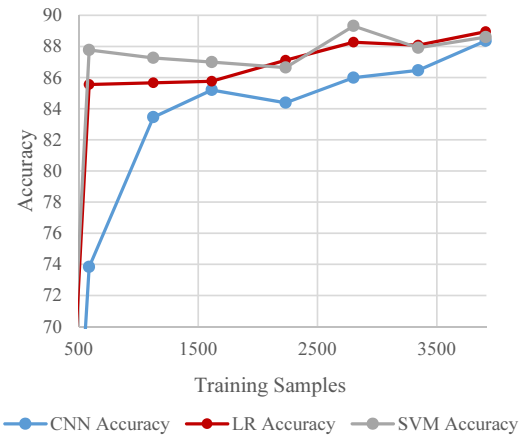

Figure 3: Improvement in test accuracy as number of training samples increases

Given a larger dataset, we believe that the CNN model and combined models involving the CNN will improve. Shown in Figure 3, CNN test accuracy improves steadily as the number of training samples increases. In contrast, classic machine learning methods such as logistic regression and SVM are less sensitive to training sample count

Discussion

Related work

Beyond the FDA MAUDE database, analysis of 48 PSE databases revealed significant technological limitations (i.e. insufficient interoperability, lack of report quality validation) that limit data entry quality. Providing embedded validators, enabling feedback review, and improving interoperability are areas of potential database improvement [21].

Substantial precedent exists in the using machine learning models to extract PSE subcategories. A hybrid CNN and bidirectional long short-term memory (LSTM) autoencoder model has been used in the detection of bleeding events from electronic health record notes with an overall F-score of 0.938 [22], and an artificial neural network (ANN) was used to identify variables which were analyzed via logistic regression to classify fall events from a nursing PSE database from a Taiwan medical center [23].

Growing and using an HIT event database

In our previous study, the models were trained on a small dataset with 490 MAUDE reports (312 HIT, 178 Non-HIT), and reached up to 88.6% accuracy, performing similarly to the classic models in this study [14]. However, we did not have an individual test set in the previous study due to the limited data size, which made the model suffer from severe overfitting. In addition, to obtain high precision, we traded off recall in the previous models. In this study, we used an upgraded dataset

from 9 years (2008-2016) of keyword-filtered MAUDE reports while the previous study only used reports from 2015 [11, 14]. We also upgraded the classifiers and proposed the SVM+LR model for HIT event identification. With an accuracy of ~90% and f1-score of ~.87, the SVM+LR+CNN model outperforms previous models and holds promise in growing an HIT event database.

To extract contextually relevant themes and topics regarding HIT events, we can use the database to process a high quantity of HIT reports at a time. Natural language processing (NLP) techniques may be the best method for identifying connections between HIT events. Such connections will provide experts with the information necessary to address the root causes of HIT events. Developing effective solutions to prevent future HIT events is the final step of this process. As HIT becomes increasingly influential in healthcare, utilization of the constantly updated HIT database will help ensure that HIT functions accordingly to ensure patient safety.

Importance of an HIT event database

The inappropriate use, design, and implementation of HIT are profoundly detrimental to patient safety. However, identifying and fixing relevant HIT issues is challenging without an organized database of HIT events. Many seemingly insignificant events may signify deeper issues, which, without a broader database, are difficult to identify. A database will allow for the detection of broader trends indiscernible from individual documents. Identifying such trends is crucial to fixing HIT issues and preventing future accidents. The fact that the nature of HIT events differs from PSEs as a whole further necessitates the accumulation of an HIT event database. For instance, contributing factors to HIT events such as equipment/device function and data availability do not exist for non-HIT events. The prevalence of certain contributing factors differs between PSE subtypes as well. While poor staff communication is a contributing factor in an estimated 40.0% of HIT events, it only factors into 24.8% of non-HIT events [24]. The uniqueness and impact of HIT mandates the HIT event database construction.

Probability outputs as a reporting aid

Screening the quality of new entries into PSE databases is another means of improving event reporting. Our model can interpret the unstructured text of PSE reports during reporting and continually update the reporter on the probability that the report is HIT or non-HIT. This yields numerous benefits. First, if the reporter writes a report description but has difficulty categorizing the report, our model can help inform their decision on whether or not to check the HIT-related box. Conversely, if the reporter erroneously labels a report, probability predictions may better inform the reporter on what label the report should be. This functionality will help reporters avoid labeling/entry mistakes. Finally, if the reporter is certain that the PSE is clearly HIT or non-HIT but unsure of how to write the report, the model's probability prediction could help guide the reporter into writing narrative text that best coincides with the report category. The overarching benefits of improved report descriptions, fewer mistakes, and more accurate labeling are that HIT-event extraction will be easier in the future, and more high-quality data will be available for analysis.

Tracking HIT changes

In 2015 alone, the U.S. Patent and Trademark Office granted 17,596 patents for new medical devices [25]. HIT is constantly changing, and we must adjust our model to captures such changes. As new PSE and HIT events are added to the MAUDE database, we can update our model. Alternatively, we could completely retrain our model on new data each year to avoid incorporating out of date HIT events. Meticulously choosing a training set that balances the inclusion new HIT events and exclusion of irrelevant, old HIT events is key to constructing a model reflective of the present nature of HIT.

Challenges of identifying HIT events with machine learning

NLP is limited in its capability to account for word context, sentence meaning, and typographical errors [26]. Generalized NLP and GloVe word embedding provide substantial information on individual word frequency and meaning, but cannot discern differences due to word context. Numerous factors such as individual writing style, interactions between different document sections, and accidental errors cannot be accounted for by NLP and machine learning models. Manual review may remain a useful supplement in the classification of reports poorly identified by NLP and machine learning.

Limitations

Missing, duplicated, and non-standardized entries were common in the MAUDE database. Out of 1,103 medical device reports, 64 reports with duplicated information and 6 reports referring to multiple patients were found. The investigation also identified the varying quality and completeness of reports is another issue [27]. Despite its flaws, the MAUDE database is the only publicly accessible database suitable for the extraction of HIT events.

Future work

Optimizing and improving our model

Deducing the quantity and precision of HIT reports desired for database construction is crucial. Because our database must be HIT-exclusive, false positives are more unfavorable than false negatives. Precision takes precedence over recall, and increasing the threshold needed to label an event as HIT beyond 50% without significantly reducing recall would be optimal.

Other deep learning models may perform better than classic models or the CNN. We will apply a recurrent neural network (RNN) for HIT event classification. Bidirection, hierarchical models, and attention may also improve RNN performance [28]. Recurrent CNN (RCNN) also offers potential for HIT event classification [29].

Application to other categories of PSEs

While PSE databases already exist for certain PSE subtypes such as medication error, two additional PSE subtypes that lack a specific database are patient falls and administrative error [30]. Applying a keyword filter to the MAUDE database can help isolate training data for database generation and probability outputs with the proposed SVM+LR+CNN model.

Conclusion

Our strategy to extract HIT events from MAUDE reports for an HIT event database construction bests prior models by all significant performance metrics. Database generation is an essential first step in identifying themes, causes, and solutions to HIT events. Moreover, our model has great potential in improving the quality and accuracy of HIT event reporting.

Acknowledgements

This project is supported by UTHealth Innovation for Cancer Prevention Research Training Program Post-Doctoral Fellowship and Summer Undergraduate Fellowship (Cancer Prevention and Research Institute of Texas grant # RP160015),

and by the Agency for Healthcare Research & Quality (1R01HS022895). The Titan Xp GPU used for this research was donated by the NVIDIA Corporation.

References

[1] Patient Safety Event, in, EMS Voluntary Event Notification Tool (E.V.E.N.T.), 2018.
[2] J.G. Anderson, K. Abrahamson, Your Health Care May Kill You : Medical Errors, *Studies in Health Technology and Informatics* **234** (2017), 13-17.
[3] L.T. Kohn, J.M. Korrigan, M.S. Donaldson, To err is human: building a safer health system, in, U.S. Institute of Medicine, November 1999.
[4] C. Macrae, The problem with incident reporting, *The British Medical Journal Quality Safety* **25** (2016), 71-75.
[5] B. Chaudhry, J. Wang, S. Wu, M. Maglione, W. Mojica, E. Roth, S. C. Morton, P. G. Shekelle, Systematic Review : Impact of Health Information Technology on Quality, Efficiency, and Costs of Medical Care, *Annals of Internal Medicine* **144** (2006), 742-752
[6] P.F. Cipriano, K. Bowles, M. Dailey, P. Dykes, G. Lamb, M. Naylor, The importance of health information technology in care coordination and transitional care, *Nursing Outlook* **61** (2013), 475-489
[7] Top 10 health technology hazards for 2015, in, Emergency Care Research Institute (ECRI), November 2014.
[8] K.C. Cheung, W. van der Veen, M. L. Bouvy, M. Wensing, P.M. van der Bemt, P.A. de Smet, Classification of medication incidents associated with information technology, *Journal of American Medical Informatics* **21** (2014), 63-70.
[9] Common Formats – Scope and Reporting, in, Agency for Healthcare Research and Quality.
[10] Manufacturer and User Facility Device Experience Database (MAUDE), in, U.S. Food & Drug Administration, June 2018.
[11] H. Kang, J. Wang, B. Yao, S. Zhou, Y. Gong, Toward Safer Health Care : A Review Strategy of FDA Medical Device Adverse Event Database to Identify and Categorize Health Information Technology (HIT) Related Events. *JAMIA Open (submitted)* (2018).
[12] F. Magrabi, M.S. Ong, W. Runciman, and E. Coiera, Using FDA reports to inform a classification for health information technology safety problems, *Journal of American Medical Informatics Association* **19** (2012), 45-53.
[13] N. Japkowicz, Learning from Imbalanced Data Sets, in, AAAI Technical Report, 2000, pp. 10-15
[14] H. Kang, Z. Yu, Y. Gong, Initializing and Growing a Database of Health Information Technology (HIT) by using TF-IDF and Biterm Topic Modeling, in : *AMIA Annual Symposium*, San Francisco, CA, USA, 2018, 1024-1033.
[15] F. Chollet, *Deep Learning with Python*, Manning Publications, Shelter Island, NY, November 2017.
[16] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural Language Processing (Almost) from Scratch, *Journal of Machine Learning Research* **12** (2011), 2493-2537.
[17] J. Pennington, R. Socher, C.D. Manning, GloVe : Global Vectors for Word Representation, in : *Empirical Methods in Natural Language* Processing, Doha, Qatar, 2014, pp. 1532-1543.[18] F. Pedregosa, G. Varoquax, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research* **12** (2011), 2825-2830.
[19] J. Ramos, Using TF-IDF to Determine Word Relevance in Document Queries, in : *The First Instructional Conference on Machine* Learning, Piscataway, NJ, USA, 2003.
[20] Y. Kim, Convolutional Neural Networks for Sentence Classification, in: *Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014.
[21] Y. Gong, H. Kang, X. Wu, L. Hua, Enhancing Patient Safety Event Reporting, *Applied Clinical Informatics* **8** (2017), 893-909.
[22] R. Li, B. Hu, W. Liu, F. Liu, F. Cunningham, D.D. McManus, H. Yu, Detection of Bleeding Events in Electronic Health Record Notes Using Convolutional Neural Network Models Enhanced with Recurrent Neural Network Autoencoders : Deep Learning Approach, *JMIR Medical Informatics* **7** (2019).
[23] T. Lee, C. Liu, Y. Kuo, M. Mills, J. Fong, C. Hung, Application of data mining to the identification of critical factors of patient falls using a web-based reporting system, *International Journal of Medical Informatics* **80** (2011), 141-150.
[24] R. Mardon, L. Olinger, M. Szekendi, T. Williams, E. Sparnon, K. Zimmer, Health Information Technology Adverse Event Reporting: Analysis of Two Databases in, The Office of the National Coordinator for Health IT, November 25, 2014.
[25] Medical Devices : All Classified Utility Patents (OR/XR), in, U.S. Patent and Trademark Office Patent Technology Monitoring Team Report, December 31, 2015.
[26] I. Zuckerman, D. Litman, Natural Language Processing and User Modeling : Synergies and Limitations, *User Modeling and User-Adapated Interaction* **11** (2001), 129-158.
[27] J.M. Sandberg, I. Gray, A. Pearlman, R.P. Terlecki, An Evaluation of the Manufacturer and User Facility Experience database that inspired the United States Food and Drug Administration's Reclassification of transvaginal mesh, *Investigative and Clinical Urology* **59** (2018), 126-132.
[28] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical Attention Networks for Document Classification, *Conference of the North American Chapter of the Association of Computational Linguistics : Human Language Technologies*, New Orleans, Louisiana, 2016.
[29] S. Lai, L. Xu, K. Liu, J. Zhao, Recurrent Convolutional Neural Networks for Text Classification, *Twenty-Ninth AAAI Conference on Artificial Intelligence*, Austin, TX, USA, 2267-2273.
[30] Z.R. Wolf, *Patient Safety and Quality : An Evidenced-Based Handbook for Nurses*, Agency for Healthcare Research and Quality, Rockville, MD, March 2008.

Address for correspondence

Yang Gong, MD, PhD

7000 Fannin St Suite 600, Houston, TX 77030, USA

Yang.Gong@uth.tmc.edu

Phone: +1 713 500 3547, Fax: +1 713 500 3907.