Natural Language Processing to Extract Meaningful Information from Patient Experience Feedback

Khalid Nawab[1] Gretchen Ramsey[2] Richard Schreiber[3]

[1] Department of Medicine, Geisinger Holy Spirit Hospital, Camp Hill, Pennsylvania, United States
[2] Patient Experience, Geisinger Holy Spirit Hospital, Camp Hill, Pennsylvania, United States
[3] Physician Informatics and Department of Medicine, Geisinger Health System, Geisinger Commonwealth School of Medicine, Camp Hill, Pennsylvania, United States

Address for correspondence Richard Schreiber, MD, FAMIA, Physician Informatics and Department of Medicine, Geisinger Health System, Geisinger Commonwealth School of Medicine, 431 North 21st Street, Suite 101, Camp Hill, PA 17011-2204, United States (e-mail: rschreiber@geisinger.edu).

Appl Clin Inform 2020;11:242–252.

Abstract

Background Due to reimbursement tied in part to patients' perception of their care, hospitals continue to stress obtaining patient feedback and understanding it to plan interventions to improve patients' experience. We demonstrate the use of natural language processing (NLP) to extract meaningful information from patient feedback obtained through Press Ganey surveys.

Methods The first step was to standardize textual data programmatically using NLP libraries. This included correcting spelling mistakes, converting text to lowercase, and removing words that most likely did not carry useful information. Next, we converted numeric data pertaining to each category based on sentiment and care aspect into charts. We selected care aspect categories where there were more negative comments for more in-depth study. Using NLP, we made tables of most frequently appearing words, adjectives, and bigrams. Comments with frequent words/combinations underwent further study manually to understand factors contributing to negative patient feedback. We then used the positive and negative comments as the training dataset for a neural network to perform sentiment analysis on sentences obtained by splitting mixed reviews.

Results We found that most of the comments were about doctors and nurses, confirming the important role patients ascribed to these two in patient care. "Room," "discharge" and "tests and treatments" were the three categories that had more negative than positive comments. We then tabulated commonly appearing words, adjectives, and two-word combinations. We found that climate control, housekeeping and noise levels in the room, time delays in discharge paperwork, conflicting information about discharge plan, frequent blood draws, and needle sticks were major contributors to negative patient feedback. None of this information was available from numeric data alone.

Conclusion NLP is an effective tool to gain insight from raw textual patient feedback to extract meaningful information, making it a powerful tool in processing large amounts of patient feedback efficiently.

Keywords

- natural language processing
- knowledge modeling and representation
- patient satisfaction
- patient engagement
- patient
- consumer health

received
November 5, 2019
accepted after revision
February 1, 2020

DOI https://doi.org/10.1055/s-0040-1708049.
ISSN 1869-0327.

Background and Significance

Irwin Press, PhD, of the University of Notre Dame was one of the first to conceive of the concept of the patient's experience of care in the early 1980s.[1] He demonstrated that understanding nonclinical needs of the patient can improve care and reduce malpractice claims.[1] The idea gained popularity. By 1984 patient satisfaction was a hot subject for many hospital administrators, but there was a lack of means of measuring it. Dr. Press partnered with Rod Ganey, PhD, who was a renowned statistician and survey methodology specialist to develop the first survey to measure patient satisfaction scientifically and improve health care. The Press Ganey survey was the result of this collaboration.[2]

The federal government became involved with patient satisfaction metrics in 2002. The Center for Medicare and Medicaid Services (CMS) partnered with the Agency for Healthcare and Research Quality (AHRQ) and developed the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey after a rigorous scientific process.[3] The public gained the opportunity to comment later.[4]

The Deficit Reduction Act of 2005 tied financial incentives to HCAHPS reporting by hospitals.[3] HCAHPS consists of 27 standardized questions administered randomly by approved third parties or the hospital to adult patients discharged from the hospital. The National Quality Forum approved the HCAHPS survey after extensive review and public feedback in May 2005.[4] CMS then implemented it in October 2006, with voluntary reporting by hospitals March 2008.[3] The Patient Protection and Affordable Care Act of 2010 then implemented the pay for performance model under which hospital Medicare reimbursement was determined in part by comparative performance and improvement of HCAHPS.[5]

Clinical outcomes in the top HCAHPS tercile hospitals are associated with lower in-hospital mortality compared with facilities in the bottom tercile.[6] Some studies have shown that hospitals with higher patient satisfaction scores have lower readmission rates for certain medical conditions,[7] improved adherence to guidelines and improved mortality in myocardial infarction[8] and shorter inpatient stays, and lower readmission and lower mortality in surgical patients.[9] More and more users now rely on internet sources and client feedback for choosing a business or a product[10] and hospitals are no different. Health care facilities rely on survey methodologies such as Press Ganey, which uses a third party to contact patients after discharge or visit date to have them complete a survey. However, feedback alone is not enough. Understanding the information contained in the raw data is as difficult as it is important, as the patient's perception of care is a complicated affair. More research is needed to determine interventions that may improve patient's perception of their care.[11]

The Press Ganey survey responses are provided to the hospital on a regular basis by the agency collecting the surveys from the patients. This is in the form of comments in response to questions included in the survey. The questions are related to various care aspects, and each comment by the patient is labeled according to the care aspect about which the question was asked: doctor, nurse, stay/room, meal, tests and treatments, admission, discharge, labor and delivery, postpartum, visitor, personal issues, overall assessment, and general comments. The responses are also labeled based on the sentiment as positive, negative, neutral, or mixed. It is then incumbent upon the hospital to review these comments and extract information pertaining to patient experience.

Patient feedback is mostly in raw text and in natural language unlike data collected during a scientific study which is generally numerical. One method to gain in-depth understanding of the feedback is manual reading of each comment or review. This is feasible with a small amount of data, but large health systems have large patient volumes and thus collect large datasets pertaining to patient experience. Combing through these data manually would require a lot of personnel resources and is not feasible. Furthermore, the mixed comments carry multiple sentiments and may be about more than one care aspect; therefore, it is a challenge to extract information from such comments. The advent of natural language processing (NLP) algorithms makes it far more convenient to analyze these data.

The use of NLP for extraction of meaningful information from patient feedback is not new and has been studied by various studies in the recent past. López et al applied sentiment analysis to patient feedback available online on different websites.[12] Ellimoottil et al also looked into sentiments in reviews available online regarding 500 urologists.[13] Doyle et al and Doing-Harris et al also examined various aspects of patient experience by using topic modeling.[14,15] Li et al applied a mixed methods approach involving literature review, human annotation, and NLP with machine learning-based models for topic modeling on patient reviews about doctors and their care, producing impressive results.[16] NLP has also been used for the analysis of Press Ganey data. Doing-Harris et al used NLP combined with a machine learning model to analyze patient feedback from Press Ganey for sentiment polarity. They also used a similar approach for topic modeling to extract patient feedback regarding certain aspects of their care. They were able to identify unexpected aspects in negative feedback such as appointment access using such a model.[15]

Unlike Doing-Harris's approach, we focused on inpatient satisfaction scores. In addition, we looked at positive as well as negative sentiments. We used an open access python library[17] to train the sentiment dictionary, rather than develop our own, which Doing-Harris et al did with their vocabulary-based and Naïve Bayes' classifiers.

Objective

The main objective of this analysis was to build on the prior work done on patient feedback using NLP, and to perform a comprehensive analysis of multiple aspects of patient feedback. We hypothesized that free-texted comments in our patient experience surveys could provide valuable information beyond the quantitative data. Since free text data are difficult to mine, we further hypothesized that NLP might offer a method to gain deeper insight into comments and patients' perceived experiences. Finally, we hypothesized that open access python programming would suffice for this study as it offers user-

friendly data structures, a large standard library, is relatively easy to learn, and is freely available. These characteristics make this method appealing to others who may wish to emulate our project. Furthermore, we demonstrate that the preclassified sentiment-based data from these surveys can be used as a dataset to train a deep learning text classification model to extract sentiment and care aspect information from mixed sentiment comments.

Methods

The Geisinger Holy Spirit in Camp Hill, Pennsylvania Department of Patient Experience provided all 2,830 patient experience surveys, including comments, submitted between January 1, 2018 and January 25, 2019 from the Press Ganey database of randomly selected patients discharged from the hospital. No patient names were visible to the researchers. Names of employees occasionally appeared in some of the comments but were removed during preprocessing (see below). This database constituted the raw data for subsequent analysis.

Preprocessing of Data

Preprocessing of textual data is the first and an important step in processing of text that has been proven to improve performance of text/document classification models.[18] The goal of preprocessing is to "standardize" the text. Preprocessing included conversion of all text to lowercase since in an algorithm "Doctor" and "doctor" are different words. This process also removes special characters and numbers. A separate process called "stemming" then reduces different forms of the same word to a common base form.[19] Considering these comments were raw text from patients, certain words were common such as "doc," "dr," and "doctor." Therefore, one term (e.g., "doctor") substituted for such synonyms. This limits the resulting number of terms to prevent redundancy, which would result in lower statistical validity. Certain words do not add any meaning to the text and these are called "stop words."[19] We also performed stop word elimination to simplify the dataset. The process is visualized in ►Fig. 1.

Data Analysis

We started the analysis by making charts to display trends that may provide meaningful information. We displayed frequently appearing words as "word cloud" in positive as well as negative comments to assess for frequency represented by font size of each word in the comments.

We displayed the total number of comments in each category based on the sentiment as a pie chart. The number of positive and negative comments in each care aspect category is also displayed as a pie chart.

In the next step, we made tables of the most common words in each category. Furthermore, the 10 most common "bigrams" were also displayed as tables. Bigrams are word pairs comprised two consecutive words as they appear in a sentence. For example, a comment "I liked my food" can be three unique bigrams: "I liked," "liked my," and "my food."

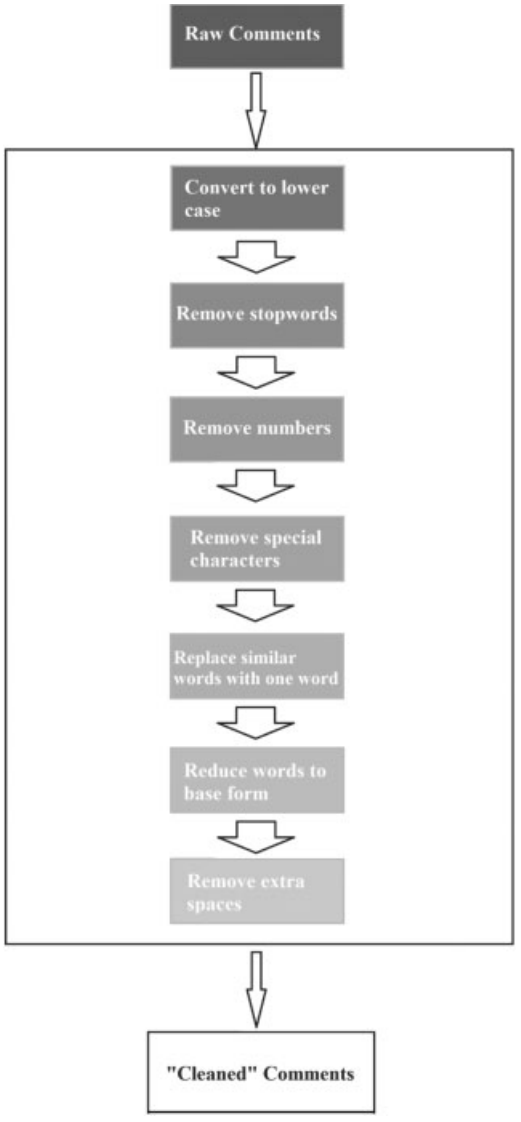

Fig. 1 Preprocessing of the raw text. All text converted to lowercase. Numbers, special characters, and extra spaces removed. Certain words were similar in meaning, for example doc, dr, and doctors, all three were replaced with physician. Words were also reduced to base form, for example explained, explaining, and explain mean the same thing, therefore reduced to one word.

"Part of speech tagging" (POS tagging) can be useful to understand how clients describe the subject matter. It is the process of assigning contextually appropriate grammatical descriptors to words in text,[20] as described by ►Fig. 2A and B.

Extracting adjectives from comments about meals may tell us how most patients felt about their meals. Similarly, extracting nouns from the same comments may help identify specific food items. Another function, "frequency distribution," tabulates words and their frequency in the text. Combining POS tagging with frequency distribution reveals the frequency of adjectives or nouns appearing in the comments.

A whole portion of the comments were mixed sentiment comments. These are the comments that carry mixed sentiments regarding more than one aspect. For example:

"The nurse was very attentive, she took very good care of me. I did not like my doctor as he did not spend much time with me."

Words	The	chicken	was	very	cold
Tags	Determiner	Noun	Verb	Adverb	Adjective

A

```
[('the', 'DET'),
 ('chicken', 'NOUN'),
 ('was', 'VERB'),
 ('very', 'ADV'),
 ('cold', 'ADJ')]
```

B

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 27)                50787
_________________________________________________________________
dropout_1 (Dropout)          (None, 27)                0
_________________________________________________________________
dense_2 (Dense)              (None, 9)                 252
_________________________________________________________________
dropout_2 (Dropout)          (None, 9)                 0
_________________________________________________________________
dense_3 (Dense)              (None, 3)                 30
=================================================================
Total params: 51,069
Trainable params: 51,069
Non-trainable params: 0
_________________________________________________________________
```

C

	precision	recall	f1-score	support
0	0.72	0.84	0.78	25
1	0.79	0.85	0.81	39
2	0.94	0.77	0.85	39
accuracy			0.82	103
macro avg	0.82	0.82	0.81	103
weighted avg	0.83	0.82	0.82	103

D

Fig. 2 Natural language processing methods: (**A**) part of speech tagging, using representation of words and their tags; (**B**) output of the parts of speech tagger. Each word is coupled with its tag; (**C**) summary of neural network based model; and (**D**) evaluation metrics of the model.

The above comment carries a positive comment toward the nurse but negative sentiment toward the doctor. Such a comment will be classified as "mixed" based on its sentiment. When reviewed by a human, very useful information can be obtained from such comments, but they provide a challenge for automated extraction with NLP. We split the mixed comments into sentences and then classified these sentences using the Keras open access python library[17] based on sentiment. Keras is a deep learning library that utilizes neural network models. In the literature, there is no clear definition of an artificial neural network (ANN). Haykin offers a good definition of ANN as a massively parallel combination of simple processing units which can acquire knowledge from the environment through a learning process and store the knowledge in its connections.[21] Deep learning models require preclassified data on which they are trained. The model extracts features or trends from the training set and then uses that to classify prospective data. We used the comments that were already classified by Press Ganey as positive, negative and neutral to train a model for sentiment-based classification.

We used the Keras Sequential model with three "dense" layers and two "dropout" layers between the dense layers. Dense layers are fully connected layers in which each neuron is connected to the neurons of the next layer. A dropout layer cancels randomly selected input neurons, and the number of neurons to be canceled is determined by a provided hyper-parameter. A summary of our model is shown in ►Fig. 2C.

The dropout layers were added to avoid overfitting. The number of neurons in the dense layers is adjusted to the best possible F1 score, as shown in ►Fig. 2D.

Different numbers of nodes were tried in the first layer and performance of the model was judged by plotting its accuracy on training and validation data as well as training and validation loss as shown in ►Fig. 3.

The data loss increased after 35 epochs; therefore, the model was stopped at that point. We applied different activation functions and found that ReLu (rectified linear unit) provided optimal accuracy. Softmax activation function was applied to the last layer of the network. This was performed on a Windows PC running with 8GB of ram and 2.4GHz Intel Core-i5 processor.

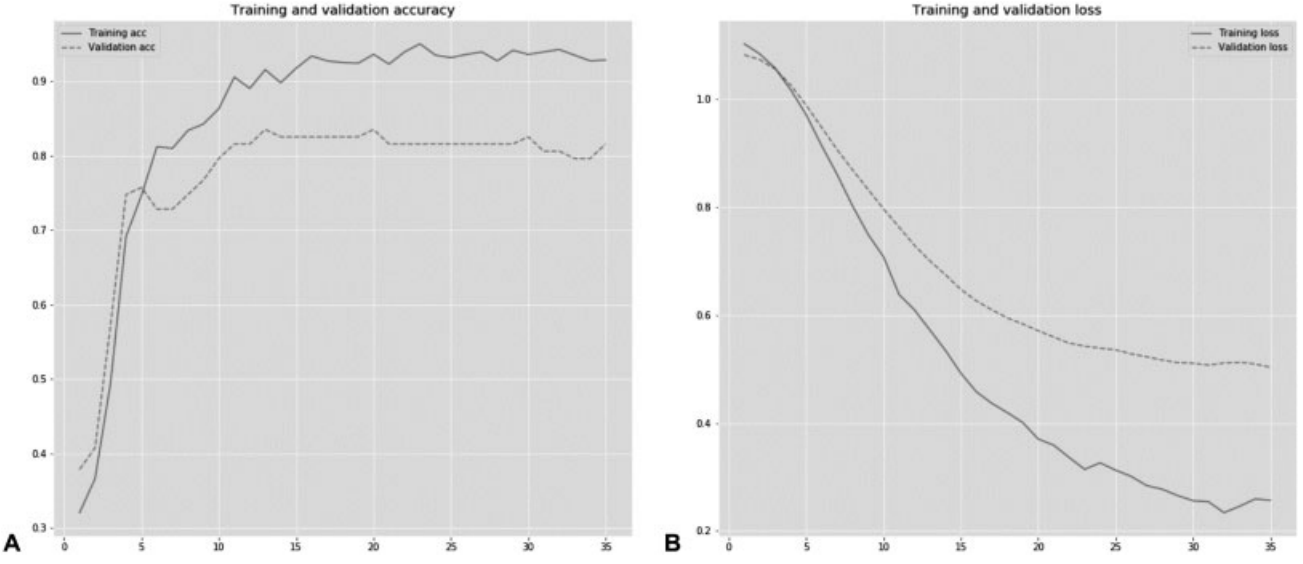

Fig. 3 (**A**,**B**) Graphical representation of the training process of the model. Accuracy did not improve significantly after 15 epochs, but loss continued to decrease and started going up after 35 epochs.

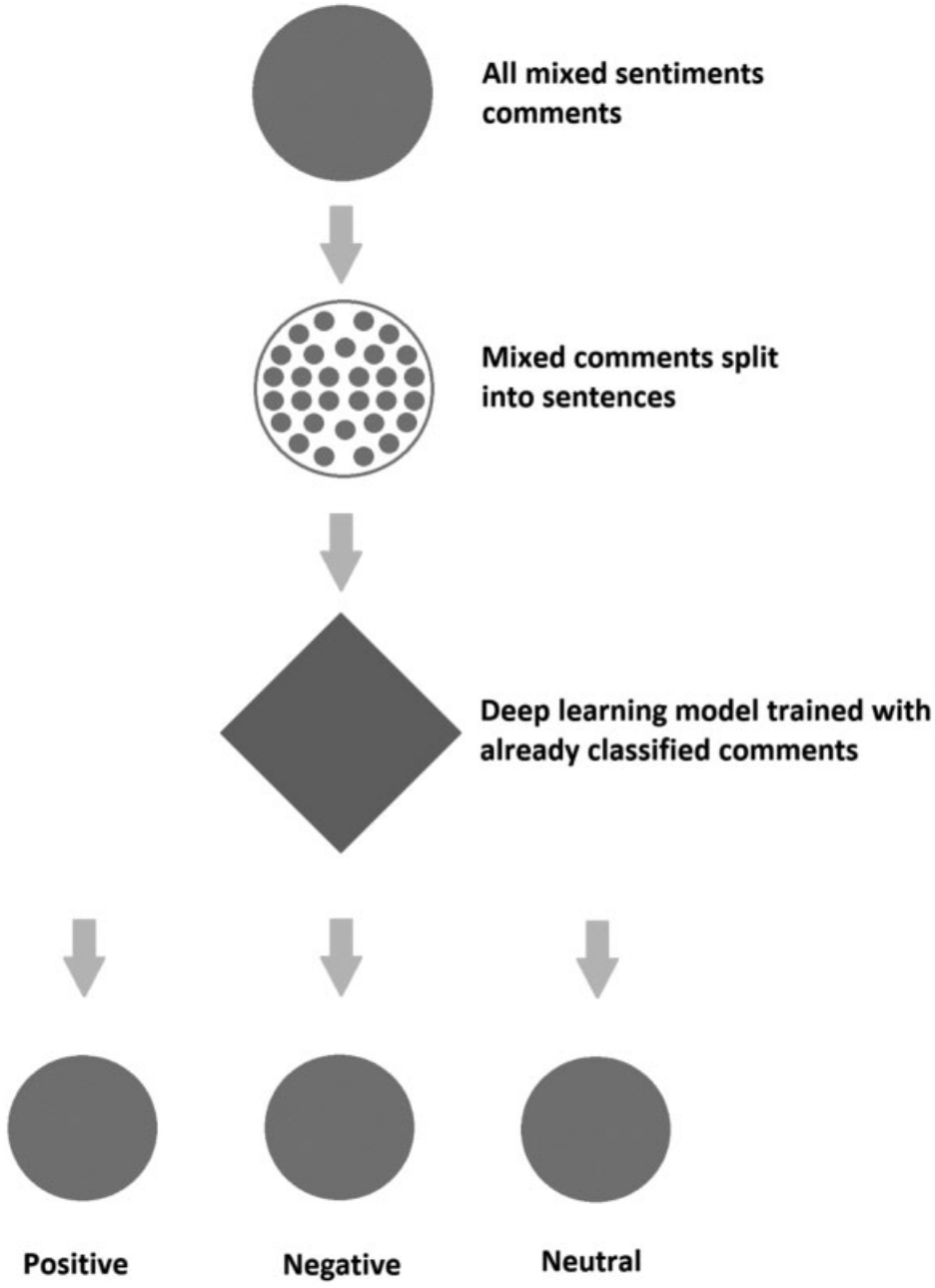

Fig. 4 Classification of mixed sentiment comments. All mixed comments were split into sentences. The deep learning model trained with the preclassified positive and negative comments was then used to classify the sentences into positive, negative, or neutral.

Once the code was written, running the model did not take more than a few minutes since the number of epochs as well as the number of neurons in the layers was relatively small.

To better explain the flow of our methodology to classify the mixed comments, refer to ►Fig. 4.

Results

►Fig. 5A reveals that "nurse" and "doctor" are the most frequent words in positive patient comments. Interestingly, the same two words appear with the highest frequency in the negative comments as seen in ►Fig. 5B. "Room" also appears frequently in the negative comments.

As shown in ►Table 1 and ►Fig. 6, there were a total of 1,332 positive comments and 849 negative comments. There were fewer and approximately equal number of mixed and neutral comments.

Based on selected aspects of care aspect, as indicated by ►Table 2 and ►Fig. 7, the largest numbers of comments were about nurses, followed by physician, supporting the central role of these professions in patient experience. The next most frequent topics were concerns about the patients' room and meals.

We compared the number of positive and negative comments in each category, as shown in ►Fig. 8. Most of the positive comments were about the nurses, followed by

Fig. 5 Word Cloud representations. The larger the word, the more frequent it appeared in the comments. (**A**) "Nurse" and "doctor" are the most frequent words in both positive and negative comments; (**B**) "room" also appears more frequently in negative comments.

Table 1 Number of comments in each category based on sentiment and care aspect

Category	Total comments	Positive	Negative	Neutral	Mixed
Admission	227	99	64	47	17
Comments	92	39	28	11	14
Discharge	239	89	91	45	14
Labor and delivery	71	51	9	2	9
Meals	289	129	82	50	28
Nurse	344	216	61	24	43
Overall	291	162	70	16	43
Personal issues	217	126	59	19	13
Physician	315	175	65	35	40
Postpartum	65	44	10	1	10
Stay	308	67	162	30	49
Tests and treatment	209	68	88	32	21
Visitors and family	162	67	60	29	6
Totals	2,829	1,332	849	341	307

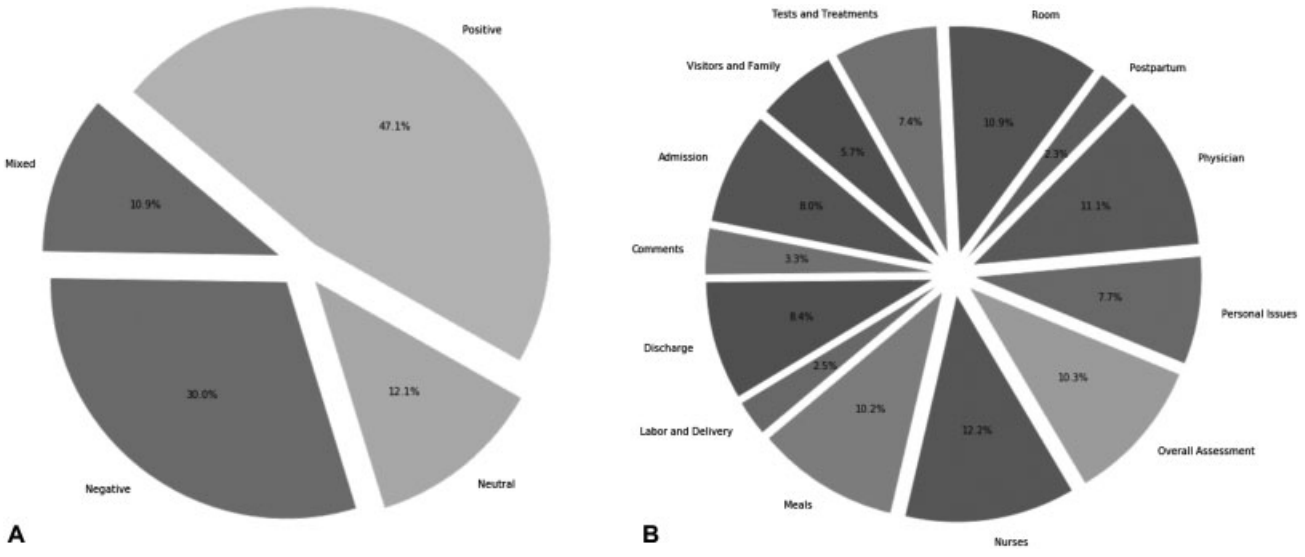

Fig. 6 Comments in each category: (**A**) based on sentiment and (**B**) based on care aspect.

Table 2 Ten most common adjectives based on selected care aspects: about hospital room, discharge, and tests and treatments

Hospital room	*n*	Discharge	*n*	Tests and treatments	*n*
Hot	18	Discharged	26	Blood	28
Small	12	Discharge	22	Intravenous	28
Loud	12	Told	20	Nurse	23
Cold	11	Doctor	19	One	22
Nurse	8	Home	18	Time	16
Noisy	7	Hour	16	Took	11
Next	5	Nurse	15	Room	11
Warm	5	Long	14	Arm	9
Uncomfortable	5	Time	14	Times	8
Clean	4	Day	12	Left	8

doctors. Most of the negative comments were about the room, followed by meals. The room had more negative comments than positive. This figure clearly shows the popularity of nurses, followed by doctors, a common theme in the comments.

An analysis of the most common adjectives used in the negative comments about the stay/room can help to understand why "room" had a significantly higher number of negative comments compared with positive comments. These are shown in ►Table 2 which represents the frequency distribution of the 10 most common negative adjectives in comments about the hospital stay. Most of the comments were about climate control ("hot," "cold," and "warm") and the size of the room ("small"). "Noisy" is another word used to describe the room. The word "next" also appeared frequently. Upon review of the comments, these were about patients in the bed next to the patient who had provided feedback, as some rooms in the facility are semiprivate.

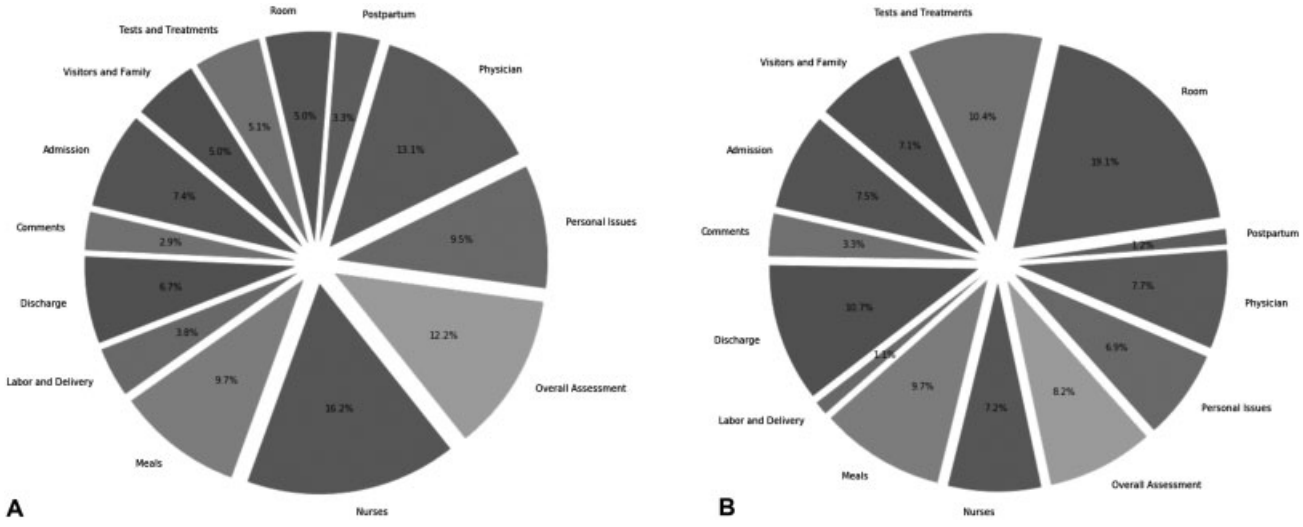

Fig. 7 Comments in each category based on care aspect: (**A**) positive comments and (**B**) negative comments.

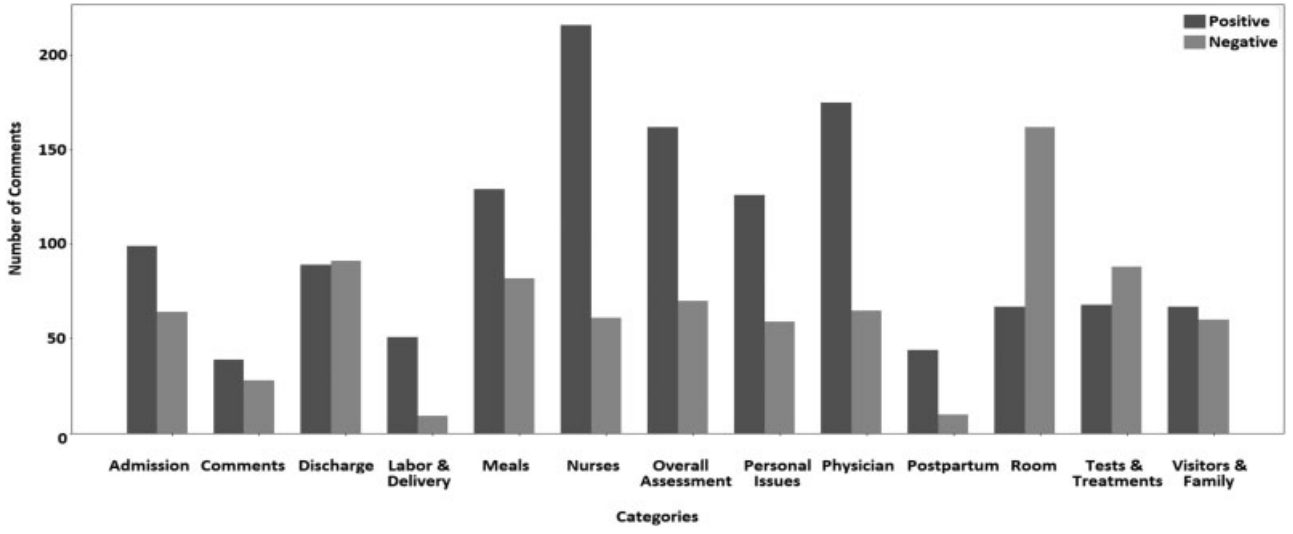

Fig. 8 Positive and negative comments in each category based on care aspect. Most care aspect categories had more positive comments than negative except for comments about room, tests and treatments, and discharge.

Table 3 Ten most common bigrams in negative comments about: discharge, and tests and treatments

Bigram: discharge	*n*	Bigram: tests and treatments	*n*
Go home	7	Took blood	5
Long time	6	AM	5
PM	5	Long time	4
Told go	4	One time	4
Took long	4	First intravenous	3
Discharge process	4	Bad experience	3
Wait hour	3	X-ray	3
Next day	3	Left arm	3
Told u	3	Back room	3
Wheel chair	3	One nurse	3

►Table 3 shows the 10 most common bigrams in negative comments about the room. As observed previously, climate control was a major concern. Bigram analysis suggests that housekeeping is also significant concern.

Another category with more negative compared with positive comments is "tests and treatments," also shown in ►Table 3. The most common word is "blood" that refers to blood draws followed by "IV," which refers to attempts of placing an IV-line access. A similar theme is seen in the bigrams as "took blood" is the most common. The rest are about the time taken by the nurse to administer a certain medication/treatment as demonstrated by "one nurse" and "long time."

The next category with more negative than positive remarks is "discharge." ►Table 3 displays that most of these comments were about the time it took to discharge the patient and late discharge. The word "told" appeared in most of these comments and upon review seemed to mean the patient was told that they were ready to be discharged, but the actual event did not occur when the patient expected.

Table 4 Mixed comments split into sentences and then classified by the deep learning model based on sentiment

All mixed comments	Total sentences	Positive	Negative	Neutral
307	895	271	297	327

Table 5 Most commonly appearing words, noun words, and bigrams in the mixed comment classified as negative by our neural model. Words that did not carry any information (e.g., "one" and "good") are not included in this table

Words	Noun	Bigrams
Nurse	Room	Th floor
Time	Time	Another nurse
Patient	Nurse	Waited long
Told	Night	Long time
Doctor	Doctor	Recovery room
Night	Floor	Night nurse

Next, we worked on extracting information from the mixed comments. There were a total of 307 comments classified as mixed based on sentiment. After splitting them into sentences, we got 895 sentences, as shown in ►Table 4.

The model does not take into account the context of the sentence and treats each sentence as a separate comment. Therefore, sentences describing events but not containing words associated with positive or negative sentiments were classified as neutral, as shown in ►Fig. 9.

A mixed comment may have many sentences. This figure shows how different sentences in one mixed feedback may be classified based on sentiment by the classification model.

However, there were a significant number of sentences classified as neutral that could have been classified as positive or negative. Most if not all of the comments classified as positive or negative did carry the sentiment assigned to it by the model. The model can be used to classify comments into aspect categories as well; however, we decided not to do so due to limited number of comments available in each category for training the model as shown in ►Table 2. Alternatively, we can get some idea about what most of the negative comments were about by looking at frequent words and word combinations as shown in ►Table 5, using our neural model.[21]

As was evident from previous comments, in the mixed comments as well, most of the negative comments were about nurse and room. The bigrams also made it evident that most of them were about long wait times. Upon review, most of these comments were regarding the time taken by the nurse to respond to the call bell. A significant number was also about the wait time for discharge papers.

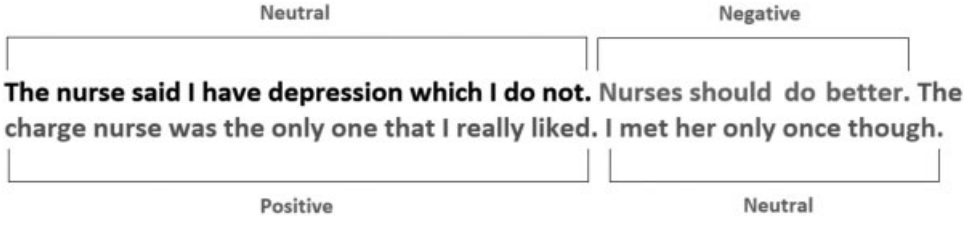

Fig. 9 Classification of mixed sentiment feedback by deep learning model. A mixed comment may have many sentences; this figure shows how different sentences in one mixed feedback may be classified based on sentiment by the classification model.

Discussion

Combining basic data plotting of numerical and textual content derived from patient feedback using basic NLP methodologies, we were able to demonstrate that large amounts of patient comments can be processed efficiently to obtain meaningful and actionable information. We found that patient comments focus on nurses and doctors who play a major role in patient experience. These two terms appeared in the highest frequency in both positive and negative comments.

Most of the negative comments regarded aspects of the patient's stay. Frequency distributions of the most common words and bigrams showed that most of these comments were about climate control, housekeeping, and noise levels. There were also more negative than positive comments about tests and treatments. A lot of the negative feedback concerned blood draws and IV sticks/lines. A significant number of negative comments concerned the time taken to discharge the patients. This included time for paperwork after patients were told they were being discharged, and one provider or nurse telling the patient they were cleared for discharge but another provider did not agree.

We demonstrated that NLP enhances the analysis of patient feedback by discovering words or combination of words appearing most frequently in the comments which provides important information about factors contributing to patient experience and which may be inaccessible from numerical analysis alone. Word combinations (bigrams, trigrams, or n-grams) in NLP may bring to light an aspect that is not obvious when looking at just most common words. The combinations of certain adjectives with both the patient's room and the food illustrate that the same adjectives may be either positive or negative sentiments depending on context. Similarly, combining analysis of bigrams with common word frequencies as well as with care aspects reveals more nuance and insight than with assessment of simple frequencies alone. Simple plotting of the number of positive and negative comments in each category into charts as shown in ►Fig. 9 identified the categories that were contributing to overall negative feedback from patients. Using NLP we identified patterns in the negative comments and on manual review of selected comments we were able to identify factors contributing to the negative feedback.

Various tools have been evaluated for understanding feedback regarding health care, but most of these tools rely on health care-related comments available on social media and have not been compared with systems working on health care-related corpus.[22] NLP has emerged as an important tool for processing unstructured clinical free text and generating structured output.[23] However, most of the work has been done on extracting healthcare-associated information from social media.[24,25] Doing-Harris et al[15] showed how NLP can be applied to discover unexpected topics in negative patient comments obtained from Press Ganey database. They used Naïve Bayes for text classification, while we implemented a neural network. Neural networks have been shown to perform better than traditional machine learning models[26,27] for text classification, in addition to other tasks like image classification. Rather than relying on the topics provided by Press Ganey, Doing-Harris et al[15] performed automatic topic modeling. This approach is very effective as it does not limit information to specific topics and lets the model discover unexpected topics.

The feedback comments provided by Press Ganey are already classified into aspect categories; therefore, one may question the utility of NLP. Even though preclassified, understanding what the negative comments are specifically talking about still requires reading through the comments. NLP makes this process efficient by identifying trends in the comment. For example, in our study, we could see that climate control was a common complaint among patients. Therefore, without reading through the comments, we were able to recognize that using NLP. A significant number of comments is also classified as mixed by Press Ganey. These comments contain mixed sentiments regarding various care aspects. NLP combined with deep learning/machine learning can extract information regarding specific aspects in these comments.

Limitations

There are certain limitations to using NLP for patient experience deep learning. For these algorithms to give meaningful information, it is very important to preprocess the data. As such, the computer does not understand the difference between words like "doctor" and "physician," even though both have the same meaning. Similarly, "meal" and "meals" carry the same information, but the computer will consider them as two different words. This extends to spelling mistakes as well where missing or replacement of a single letter causes the computer to consider it as a completely different word. The processes of stemming or lemmatizing can overcome this by reducing words to their base form or to a common word. NLP libraries are also available to correct spelling mistakes. A one-time manual analysis of the comments can identify words specific to the context and subsequent manual coding can replace them with a common word. The amount of preprocessing required for patient-derived data are also more complex since respondents may lack health care literacy, may make spelling mistakes, or use words common to texting but which may not be a part of the NLP library used for such processing.

Using frequency distribution of words or combinations of words can provide meaningful information, but it is very important to understand the context of these words which may be easy to miss while looking at tables, especially for isolated words. Thus, even though useful information in the numeric tables, one may still have to read through the feedback to understand the context. Such tables are still useful in highlighting frequently appearing words and thus provide some direction regarding effective use of patient feedback.

For a limited amount of data, NLP may not be very expedient. Techniques such as frequency distributions, bigrams, and trigrams rely on repetitive words and word combinations and with a smaller number of comments, the results may not be as fruitful and there may not be enough raw data to detect a specific pattern. On the other hand, NLP offers an advantage for

large datasets as it provides the sophisticated tools to understand it. By identifying patterns, NLP processes can also direct the investigators and identify comments that may carry more useful information and help to decrease manual analysis.

The volume of data representing patient or client feedback is increasing daily as is evident from the numerous questionnaires, third-party surveys, and internet reviews which are openly available to the public. Understanding these data plays a major role in devising interventions to improve patient experience while increasing the need for greater resources to understand the feedback. This is where NLP plays an important role as it can process large amounts of data efficiently. Combining machine learning models with the powers of NLP can help train those models to classify comments into certain categories, split mixed comments into separate single comments for better insight, and identify words that are different but used in similar contexts and thus impart similar meaning. Our model treated each sentence obtained from splitting a mixed sentiment review as an individual comment, in other words, it did not take into consideration the context of each sentence. Using a "context aware" algorithm/model, such as BERT,[28] will certainly improve the classification accuracy of these models and provide more useful information. Health care organization can certainly benefit from such systems as they deal with data volume, velocity, and variety inundation.

Conclusion

Using NLP, we were able to identify major contributors to negative patient experiences in an efficient way that did not require reading through all the comments, which then fosters a greater ability to devise solutions to these issues. NLP may not replace analysis of patient feedback by humans but will certainly make the process more efficient to seek comments which likely carry more valuable information as well as eliminating text from the mixed comments that do not carry any sentiment. Furthermore, preclassified data provided by Press Ganey can be used as a training set for a machine learning or deep learning model that can effectively and prospectively classify these data into actionable information for health care institutions. Our study demonstrates that free-texted feedback obtained from patient experience surveys can be analyzed efficiently and effectively using NLP and potential important information pertaining to patient experience can be extracted from it without having to manually read through all the feedback.

Clinical Relevance Statement

Understanding the patient experience not only enhances patient care, but helps understand patient concerns including areas where a health care organization is performing well and areas where opportunities for improvement exist. For more than a decade, health care organizations have been using quantitative data provided in surveys, but free-texted comments are also available. We leveraged natural language programming to provide in-depth analysis of free-texted comments in patient experience surveys. We recommend that other organizations take advantage of these tools. One implementation would be an NLP-based tool embedded into the system for patient experience department that is fed with Press Ganey feedback comments on ongoing basis to extract important information about factors contributing to patients' perception of their care, allowing the department to extract information without the need to go through large number of patient comments.

Multiple Choice Questions

1. Free-texted comments by patients produce many synonyms for the same concept. NLP can reduce these sources of redundancy by a process called:
   a. Creation of bigrams
   b. Creation of word clouds
   c. Lemmatizing
   d. Part of speech tagging

   Correct Answer: The correct answer is option c lemmatizing which also called stemming. Using only the stem of a word allows for grouping of conceptually similar words without regard to part of speech, spelling, abbreviations, or other redundancies. Bigrams are pairs of words and do not correct for such synonyms. A word cloud provides a frequency display of words, bigrams, trigrams, and other terms. Part of speech tagging identifies words as noun, verb, and so forth, which allows for frequency analysis, concept clarification, and categorization into aspects of care.

2. This study revealed that of the more than 2,000 comments, patients offered the most positive and negative comments regarding:
   a. Doctors, blood draws, and intravenous
   b. Doctors, nurses, and discharge
   c. Doctors, nurses, and the food
   d. Doctors, nurses, and the room

   Correct Answer: Although there were many comments about blood draws, intravenous, and the discharge process, overall the most frequent comments concerned nurses, followed by doctors, followed by comments about the room.

Authors' Contributions

K.N. and G.R. conceived of and developed the project. G.R. provided and analyzed the Press Ganey data. K.N. performed the programming and analysis of the data and wrote the first draft of the manuscript. R.S. provided substantial support including encouragement to publish, as well as extensive writing and editing of the manuscript. All authors approved the final manuscript for submission.

Protection of Human and Animal Subjects

The Geisinger Health System Institutional Review Board ruled that this project was not subject to its oversight as the proposal is "research that does not involve human subjects" as defined in 45 CFR 46.102(f).

Funding
None.

Conflict of Interest
None declared.

References

1 Press I. Concern for the patient's experience comes of age. Pat Exper J 2014;1(01):4–6
2 Press Ganey. History & Mission. Available at: https://www.pressganey.com/about/history-mission. Accessed December 11, 2019
3 CMS.gov. HCAHPS: Patients' Perspectives of Care Survey. Available at: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalHCAHPS. Accessed December 11, 2019
4 CMS.gov. The HCAHPS survey–Frequently asked questions. Available at: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/HospitalHCAHPSFactSheet201007.pdf. Accessed December 11, 2019
5 Office of the Legislative Counsel. Patient Protection and Affordable Care Act; Health-related portions fo the Heatlh Care and Education Reconciliation Act of 2010. Available at: http://housedocs.house.gov/energycommerce/ppacacon.pdf. Accessed December 11, 2019
6 Dottino JA, He W, Sun CC, et al. Centers for Medicare and Medicaid Services' Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) scores and gynecologic oncology surgical outcomes. Gynecol Oncol 2019;154(02):405–410
7 Schron E, Friedmann E, Thomas SA. Does health-related quality of life predict hospitalization or mortality in patients with atrial fibrillation? J Cardiovasc Electrophysiol 2014;25(01):23–28
8 Dominick KL, Ahern FM, Gold CH, Heller DA. Relationship of health-related quality of life to health care utilization and mortality among older adults. Aging Clin Exp Res 2002;14(06):499–508
9 Idler EL, Benyamini Y. Self-rated health and mortality: a review of twenty-seven community studies. J Health Soc Behav 1997;38 (01):21–37
10 Maslowska E, Malthouse EC, Viswanathan V. Do customer reviews drive purchase decisions? The moderating roles of review exposure and price. Decis Support Syst 2017;98:1–9
11 Davidson KW, Shaffer J, Ye S, et al. Interventions to improve hospital patient satisfaction with healthcare providers and systems: a systematic review. BMJ Qual Saf 2017;26(07):596–606
12 López A, Detz A, Ratanawongsa N, Sarkar U. What patients say about their doctors online: a qualitative content analysis. J Gen Intern Med 2012;27(06):685–692
13 Ellimoottil C, Hart A, Greco K, Quek ML, Farooq A. Online reviews of 500 urologists. J Urol 2013;189(06):2269–2273
14 Doyle C, Lennox L, Bell D. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. BMJ Open 2013;3(01):e001570
15 Doing-Harris K, Mowery DL, Daniels C, Chapman WW, Conway M. Understanding patient satisfaction with received healthcare services: a natural language processing approach. AMIA Annu Symp Proc 2016;2016:524–533
16 Li J, Liu M, Li X, Liu X, Liu J. Developing embedded taxonomy and mining patients' interests from web-based physician reviews: mixed-methods approach. J Med Internet Res 2018;20(08): e254
17 Keras: The Python Deep Learning library. Available at: https://keras.io. Accessed December 11, 2019
18 Chandrasekar P, Qian K. The Impact of Data Preprocessing on the Performance of a Naïve Bayes Classifier. In: Proceedings - International Computer Software and Applications Conference. Vol 2. IEEE Computer Society2016:618–619. Doi: 10.1109/COMPSAC.2016.205 Available at: https://www.semanticscholar.org/paper/The-Impact-of-Data-Preprocessing-on-the-Performance-Chandrasekar-Qian/f624888aa484238383513a406accb2a958ed90d9. Accessed February 18, 2020
19 Vijayarani S. Research scholar MP. Preprocessing techniques for text mining-an overview. Int J Comp Sci Comm Networks 2015;5 (01):7–16
20 Mitkov R. The Oxford Handbook of Computational Linguistics. 1st ed. Oxford, United Kingdom: Oxford University Press; 2003. Doi: 10.1093/oxfordhb/9780199276349.001.0001
21 Guresen E, Kayakutlu G. Definition of Artificial Neural Networks with comparison to other networks. Procedia Comput Sci 2011; 3:426–433
22 Gohil S, Vuik S, Darzi A. Sentiment analysis of health care tweets: review of the methods used. JMIR Public Health Surveill 2018;4 (02):e43
23 Kreimeyer K, Foster M, Pandey A, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J Biomed Inform 2017;73:14–29
24 Ranard BL, Werner RM, Antanavicius T, et al. Yelp reviews of hospital care can supplement and inform traditional surveys of the patient experience of care. Health Aff (Millwood) 2016;35 (04):697–705
25 Nikfarjam A, Ransohoff JD, Callahan A, et al. Early detection of adverse drug reactions in social health networks: a natural language processing pipeline for signal detection. JMIR Public Health Surveill 2019;5(02):e11264
26 Parwez MA, Abulaish M. Jahiruddin. Multi-label classification of microblogging texts using convolution neural network. IEEE Access 2019;7:68678–68691
27 Lee SH, Levin D, Finley PD, Heilig CM. Chief complaint classification with recurrent neural networks. J Biomed Inform 2019; 93:103158
28 Google AI. Blog: open sourcing BERT: state-of-the-art pre-training for natural language processing. Available at: https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html. Accessed December 11, 2019