Use of Sentiment Analysis for capturing Patient Experience from reviews posted online

By
VIKNESHWAR E,
201405610
M.Tech CSE (PG2),
IIIT Hyderabad

Under the guidance of
Dr Vikram Pudi,
IIIT Hyderabad.

Abstract :

There are lot of unstructured text comments or reviews posted online about the quality of healthcare being provided. Patients comment about hospitality, quality of treatments and physicians and also provide their ratings. However these comments and ratings are not captured in a systematic way. To understand their concerns and thus to improve the quality of healthcare, we need analytical techniques such as Sentiment Analysis which identifies the polarity of the statements.

Objective :

Machine Learning Techniques are used to learn about Patients comments. Those comments are categorized as positive or negative using Sentiment Analysis Techniques. Hence the objective is to compare our observations against the ratings provided by patients.

Introduction :

Understanding patients’ experience of health care is central to the process of providing care and is a fundamental pillar of health care quality.Traditional measures of patient experience include surveys and structured patient reported outcome measures. Such approaches ask specific and limited questions. Surveys are conducted infrequently, and are often expensive to administrator. Nowadays patients post their reviews on Internet in blogs,social media and on health care rating websites. However these reviews are largely unstructured and are not captured in a systematic way.This represents a missed opportunity for understanding patients’ experience in an increasingly “connected” world.

Natural Language Processing of large datasets including Sentiment Analysis Techniques and Opinion Mining Techniques have been critical to understand Consumer attributes and behaviors. For example,Sentiment Analysis Techniques are used by politicians to prepare Election Manifesto and by statisticians to predict the results of elections. If the same can be applicable to health care,these analytical methods could permit interpretation of textual information about patient experience on a huge scale. This information, because of its prose nature, has avoided the analytical spotlight of conventional quantitative analysis. Patient comments about specific doctors could be attributed with reasonable accuracy to positive and negative sentiment.

Yelp is a popular social media website that allows customers to share their business experiences with other customers.Yelp has made available an Academic Dataset of the 13,490 closest businesses to 30 universities for researchers to explore.Many methodological papers have been published on analyzing restaurants using this dataset.Yelp is a popular social media website that allows customers to share their business experiences with other customers.Yelp has made available an Academic Dataset of the 13,490 closest businesses to 30 universities for researchers to explore.Many methodological papers have been published on analyzing restaurants using this dataset.However, this data set has yet to be studied in the context of health care.

Hence reviews related to healthcare are extracted from this dataset for analysis. Natural Language Processing Techniques are used to extract the keywords that are essential to determine the sentiment score of each review. Once those keywords are identified, the feature words are converted into vector format using TFIDF Vectorizer. The feature vector is then applied to supervised machine learning algorithms(classifiers) to train them with reviews and their sentiment score. Then this classifier is used to predict the sentiment of the review which is compared against patients ratings and hence accuracy of those algorithms are determined.

Dataset Description :

Yelp dataset contains reviews and ratings provided by its users for its various business categories. Those reviews and ratings are well categorized in its dataset. This dataset can be downloaded from https://www.yelp.com/dataset_challenge . However we are interested in reviews which are related to healthcare. Hence we extracted data that are related to following 26 health care related categories.

1. Hospitals
2. General Dentistry
3. Urgent Care
4. Chiropractors
5. Health and Medical
6. Physical Therapy
7. Medical Centers

8. Eyewear and Opticians
9. Obstetricians and Gynecologists
10. Optometrists
11. Sports Medicine
12. Ophthalmologists
13. Orthopedists
14. Periodontists
15. Doctors
16. Dentists
17. Family Practice
18. Oral Surgeons
19. Allergists
20. Massage
21. Internal Medicine
22. Massage Therapy
23. Pediatricians
24. Naturopathic/Holistic
25. Laser Eye Surgery/Lasik
26. Acupuncture

Implementation :

1)Categorizing the reviews :

Data related to the above categories are extracted from yelp dataset. Patients have provided their ratings on a 5 point scale. For classification purpose,ratings that are greater than or equal to 3 have been labeled as positive and others are labeled as negative. The above dataset has both training and test data files separately . Hence the patient reviews are labeled using the above approach for both training and test data files.

However patients can have neutral opinion and can provide their ratings and reviews. Hence the dataset(which includes both training and test data files)is separated into dataset with two categories(Positive and Negative) as well as dataset with three categories(Positive,Negative and Neutral). Reviews of patients who have provided ratings as 3 are labeled Neutral while labeling the dataset for three categorizes. Hence the modified dataset contains four files Training dataset (for both two and three categories) plus Test dataset(for both two and three categories).

2)Feature Extraction :

Machine Learning Approach is in which an algorithm learns to classify comments into categories based on the given examples.A typical machine learning approach has two components 1) Data Preprocessing - Data is split into feature vectors to train the classifier based on its category 2)

Classification - Classifier uses feature vector from the Training Dataset and trains itself based on labels and fits the feature vector from Test Dataset to predict the labels for the Test Data.

Stopword removal in Sentiment Analysis is more tedious than that in Text processing. For example, “not” is a stopword that can be ignored in text processing. But while determining the polarity of a sentence,it plays a significant role. For example, The facilities that are provided here are not so great. The polarity of above statement is negative. But if “not” is removed as a stopword, this polarity will be determined as positive. Hence those Stopwords that don’t play a role in determining the sentiment of a statement only have to be removed. Hence a list of stopwords that don’t play a role in determining the sentiment of a statement is prepared. While preprocessing the user reviews, words that are present in that list are removed as they don’t have any significance in determining the sentiment of statement.

Porter Stemmer is used to stem the remaining words. For classification, we need to identify words which are similar. Example : great and greater, treat and treats etc.. Hence we apply stemming to convert those similar words into stemmed version.

Once the words are stemmed, they are converted into Feature Vectors using Bag of Words approach. The vectorizer used here is TFIDF Vectorizer.Bag Of Words is a simple unordered collection of words. For analysis, unigrams and bigrams are considered. Higher order n-grams have not been used considering the computing power and time.TFIDF Vectorizer returns feature vectors that are supposed to be passed to the classification algorithm.

3)Classification :

The feature vectors for both the training and test data files (for both two class and three class) are provided by TFIDF Vectorizer is passed to the following Classification Algorithms and their accuracies are determined. 1)Multinomial Naive Bayes 2)Support Vector Machines 3)Maximum Entropy Classifier and 4) K Nearest Neighbors(KNN) Classifier.

K Fold Cross Validation is applied to validate the performance of the classifier with respect to the training samples. Then the classifier is used to predict the labels for test samples. Performance of classifier with respect to cross validation as well as test samples are observed.

Choosing the classifier is important in determining the accuracy. Naive Bayes classifiers generally work well on text classification and also trains faster than Support Vector Machines. But Support Vector Machines can eliminate the need of feature selection. Maximum Entropy Classifier has the ability to converge to a global optimum with respect to a training set.

Classifier fits the feature vector of the training set and the class labels. Then it fits the feature vector of test set and predicts the corresponding labels. Now the predicted labels are compared against the actual labels and hence the performance of the classifier is determined. Confusion Matrices are displayed to depict the number of correctly classified and misclassified samples with respect to each class.

Observations :

1) Unigram Extraction
   i)Multinomial Naive Bayes

Cross Validation on classes with two categories(Positive and Negative)

Accuracy	Confusion Matrix
88.7	[[1301 148] [32 113]]

Cross Validation on classes with three categories (Positive,Negative and Neutral)

Accuracy	Confusion Matrix
81.36	[[1175 137 105] [19 118 13] [17 6 4]]

Final Testing on classes with two categories(Positive and Negative)

Accuracy	Confusion Matrix
87.76	[[332 45] [6 34]]

Final Testing on classes with three categories(Positive,Negative and Neutral)

Accuracy	Confusion Matrix
83.45	[[312 43 17] [5 36 4] [0 0 0]]

ii)Support Vector Machines

Cross Validation on classes with two categories(Positive and Negative)

Kernel(Parameter)	Accuracy	Confusion Matrix
RBF with C=10000000	90.46	[[1300 119] [33 142]]
Linear with C=5	90.65	[[1287 103] [46 158]]

Cross Validation on classes with three categories (Positive,Negative and Neutral)

Kernel(Parameter)	Accuracy	Confusion Matrix
RBF with C=10000000	82.93	[[1159 99 98] [27 150 11] [25 12 13]]
Linear with C=5	83.18	[[1151 91 91] [34 160 16] [26 10 15]]

Final Testing on classes with two categories(Positive and Negative)

Kernel(Parameter)	Accuracy	Confusion Matrix
RBF with C=10000000	92.36	[[323 17] [15 62]]
Linear with C=5	92.8	[[326 18] [12 61]]

Final Testing on classes with three categories(Positive,Negative and Neutral)

Kernel(Parameter)	Accuracy	Confusion Matrix
RBF with C=10000000	88.24	[[301 16 12] [14 63 5] [2 0 4]]
Linear with C=5	88.24	[[302 16 13] [14 63 5] [1 0 3]]

iii)Maximum Entropy Classifier

Cross Validation on classes with two categories(Positive and Negative)

Multi_class(Parameter)	Accuracy	Confusion Matrix
multinomial	89.39	[[1329 165] [4 96]]
ovr	87.7	[[1330 193] [3 68]]

Cross Validation on classes with three categories (Positive,Negative and Neutral)

Multi_class(Parameter)	Accuracy	Confusion Matrix
multinomial	82.18	[[1202 153 118] [8 108 4] [1 0 0]]
ovr	80.99	[[1206 176 119] [5 85 3] [0 0 0]]

Final Testing on classes with two categories(Positive and Negative)

Multi_class(Parameter)	Accuracy	Confusion Matrix
multinomial	92.08	[[335 30] [3 49]]
ovr	89.2	[[335 42] [3 37]]

Final Testing on classes with three categories(Positive,Negative and Neutral)

Multi_class(Parameter)	Accuracy	Confusion Matrix
multinomial	88.18	[[314 26 16] [3 53 5] [0 0 0]]
ovr	85.37	[[315 38 18] [2 41 3] [0 0 0]]

iv)K Nearest Neighbors

Cross Validation on classes with two categories(Positive and Negative)

K(no of neighbors)	Accuracy	Confusion Matrix
7	86.7	[[1311 190] [22 71]]
12	86.7	[[1319 198] [14 63]]
17	86.7	[[1311 190] [22 71]]

Cross Validation on classes with three categories (Positive,Negative and Neutral)

K(no of neighbors)	Accuracy	Confusion Matrix
7	79.73	[[1192 182 114] [18 79 8] [1 0 0]]
12	79.79	[[1188 177 114] [22 84 8] [1 0 0]]
17	79.48	[[1190 184 115] [21 77 7] [0 0 0]]

Final Testing on classes with two categories(Positive and Negative)

K(no of neighbors)	Accuracy	Confusion Matrix
7	86.33	[[329 48] [9 31]]
12	85.37	[[331 54] [7 25]]
17	86.33	[[329 48] [9 31]]

Final Testing on classes with three categories(Positive,Negative and Neutral)

K(no of neighbors)	Accuracy	Confusion Matrix
7	78.89	[[301 51 19] [16 28 2] [0 0 0]]
12	81.77	[[308 46 18] [9 33 3] [0 0 0]]
17	81.77	[[307 45 19] [10 34 2] [0 0 0]]

2) Bigram Extraction

i)Multinomial Naive Bayes

Cross Validation on classes with two categories(Positive and Negative)

Accuracy	Confusion Matrix
90.52	[[1270 88] [63 173]]

Cross Validation on classes with three categories (Positive,Negative and Neutral)

Accuracy	Confusion Matrix
83.31	[[1150 82 101] [42 175 18] [19 4 3]]

Final Testing on classes with two categories(Positive and Negative)

Accuracy	Confusion Matrix
87.52	[[337 51] [1 28]]

Final Testing on classes with three categories(Positive,Negative and Neutral)

Accuracy	Confusion Matrix
83.45	[[316 47 18] [1 32 3] [0 0 0]]

ii)Support Vector Machines

Cross Validation on classes with two categories(Positive and Negative)

Kernel(Parameter)	Accuracy	Confusion Matrix
RBF with C=10000000	90.71	[[1324 139] [9 122]]
Linear with C=5	90.71	[[1325 140] [8 121]]

Cross Validation on classes with three categories (Positive,Negative and Neutral)

Kernel(Parameter)	Accuracy	Confusion Matrix
RBF with C=10000000	83.43	[[1199 130 117] [11 131 5] [1 0 0]]
Linear with C=5	83.43	[[1199 130 117] [11 131 5] [1 0 0]]

Final Testing on classes with two categories(Positive and Negative)

Kernel(Parameter)	Accuracy	Confusion Matrix
RBF with C=10000000	92.8	[[327 19] [11 60]]
Linear with C=5	92.8	[[327 19] [11 60]]

Final Testing on classes with three categories(Positive,Negative and Neutral)

Kernel(Parameter)	Accuracy	Confusion Matrix
RBF with C=10000000	88.96	[[310 18 16] [7 61 5] [0 0 0]]
Linear with C=5	88.96	[[311 19 16] [6 60 5] [0 0 0]]

iii)Maximum Entropy Classifier

Cross Validation on classes with two categories(Positive and Negative)

Multi_class(Parameter)	Accuracy	Confusion Matrix
multinomial	85.75	[[1333 227] [0 34]]
ovr	84.19	[[1333 252] [0 9]]

Cross Validation on classes with three categories (Positive,Negative and Neutral)

Multi_class(Parameter)	Accuracy	Confusion Matrix
multinomial	79.17	[[1211 210 120] [0 51 2] [0 0 0]]
ovr	77.35	[[1211 239 122] [0 22 0] [0 0 0]]

Final Testing on classes with two categories(Positive and Negative)

Multi_class(Parameter)	Accuracy	Confusion Matrix
multinomial	87.76	[[335 48] [3 31]]
ovr	86.09	[[[337 57] [1 22]]

Final Testing on classes with three categories(Positive,Negative and Neutral)

Multi_class(Parameter)	Accuracy	Confusion Matrix
multinomial	84.89	[[[315 40 19] [2 39 2] [0 0 0]]
ovr	82.49	[[316 51 19] [1 28 2] [0 0 0]]

iv)K Nearest Neighbors

Cross Validation on classes with two categories(Positive and Negative)

K(no of neighbors)	Accuracy	Confusion Matrix
7	87.01	[[1283 157] [50 104]]
12	86.82	[[1298 175] [35 86]]
17	87.2	[[1297 168] [36 93]]

Cross Validation on classes with three categories (Positive,Negative and Neutral)

K(no of neighbors)	Accuracy	Confusion Matrix
7	79.73	[[1165 155 111] [45 106 11] [1 0 0]]
12	80.11	[[1174 158 106] [37 103 16] [0 0 0]]
17	80.36	[[1179 159 109] [32 102 13] [0 0 0]]

Final Testing on classes with two categories(Positive and Negative)

K(no of neighbors)	Accuracy	Confusion Matrix
7	83.93	[[318 47] [20 32]]
12	86.57	[[330 48] [8 31]]
17	86.09	[[322 42] [16 37]]

Final Testing on classes with three categories(Positive,Negative and Neutral)

K(no of neighbors)	Accuracy	Confusion Matrix
7	79.61	[[299 46 18] [18 33 3] [0 0 0]]
12	81.77	[[304 42 19] [13 37 2] [0 0 0]]
17	81.53	[[301 40 17] [16 39 4] [0 0 0]]

Results :

It is observed that Support Vector Machines proved to be the best classifier for this dataset with accuracy close to 92% while extracting features by both unigrams as well as bigrams. Support Vector Machines perform slightly better for bigrams compared to unigrams but the deviation in performance is very less. Multinomial Naive Bayes performs significantly better for bigrams than for unigrams. But Maximum Entropy Classifier performs relatively much better for unigrams than that of bigrams. Hence if features are extracted as unigrams then SVM > MaxEnt > NBM > KNN. But if features extracted are bigrams then SVM > NBM > MaxEnt > KNN. For SVM, the “C” values have to be carefully tuned for better results. For RBF Kernel ,the best C value was for 10000000. When C value was set for 100 or 1000, the accuracy was close to 77%. But once the C value was chosen to be 10000000, 92% accuracy is obtained. Similarly in Maximum Entropy classifier, the multi_class parameter is multinomial instead of ovr for better results. For better performance in Cross Validation, the class labels have to be uniformly distributed especially for Multinomial Naive Bayes. For K Nearest Neighbors, it is observed that for K=12 to 17, we obtain accuracy close to 85%.

It can also be observed that accuracy was much higher when you consider only two classes(Positive and Negative) compared to three classes(Positive,Negative and Neutral). There may be very less no of reviewers who would have rated 3. Overall for this dataset, Support Vector Machines is the recommended classifier.

Limitations :

Sentiment analysis via a machine learning approach is only as good as the learning set that is used to inform it.Online comments left without solicitation on a website are likely to have a natural selection bias towards examples of both good and bad care. It is likely that these online reviews are contributed more by those in particular demographic groups including younger and more affluent people.Irony, sarcasm, and humor frequently adopted by English speakers when talking about their care, cannot be easily detected using this process. The use of prior polarity improved the results and mitigated some colloquial phrasing, but there were difficulties understanding those that depend on context. For example, phrases that cropped up repeatedly, such as “stank of urine” or “like an angel”, could be easily characterized as negative or positive. The meaning of other frequently used phrases, however, was hard to establish without an understanding of their context. The best example of this was the phrase a “cup of tea”. It was referred to in many different comments in these data, but without knowing the context, is it impossible to allocate it a direct sentiment. ‘”They didn’t even offer me a cup of tea” is very different to “The nurse even made me a cup of tea”. Future attempts to improve a natural language processing ability for patient experience would have to develop the capacity to accurately interpret this level of context specific and idiomatic content.

References :

1)Collecting and Analyzing Patient Experiences of Health Care From Social Media.
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4526973/

2)Use of Sentiment Analysis for capturing Patient Experience from free text comments posted online
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3841376/

3) Extracting Sentiment from Healthcare Survey
https://isquared.files.wordpress.com/2015/01/extracting-sentiment-from-healthcare-survey-data.pdf