Building a Corpus for Sentiment Analysis of Operation Notes: Finding Complications and Concerns in Surgery

Jake Albert  
Department of Computer Science  
Yale University  
jacob.albert@yale.edu

Dragomir Radev  
Department of Computer Science  
Yale University  
dragomir.radev@yale.edu

Wade Schulz  
Department of Laboratory Medicine  
Yale School of Medicine  
wade.schulz@yale.edu

Abstract

This project involved the creation of an annotated corpus of operation notes written during and after surgeries performed at Yale New Haven Hospital (YNHH) for use in sentiment analysis. In order to create the corpus, an annotator tool was built as a JavaScript web application, and this tool was given to annotators who were instructed to rate phrases and sentences in the note, as well as the note as a whole, on whether they indicate that the operation was 1) complicated or not complicated, and 2) concerning or not concerning in its outcome. The resulting annotations were used to train a binary classifier on sentences in the notes using a Naive Bayes Classifier written in Python, and an LSTM Recurrent Neural Network (RNN) written in Python using Keras. This framework offers promise if trained on sets of several hundred notes.

1 Introduction

Sentiment analysis is an important area of Natural Language Processing and is used in a wide variety of contexts. The most well-known uses of sentiment analysis are in tasks that rate text along a “good to bad” polarity. For these tasks, annotated corpora such as the Cornell Movie Review Dataset, which consists of sentences labeled as positive or negative, and the Stanford Sentiment Treebank, which consists of parsed sentences from the Cornell set with subtrees labeled as positive or negative, are used in order to analyze how a subjective judge perceives a work of art, a commercial product, or a geopolitical event.

There are many reasons that sentiment analysis may be used on text in the medical domain. One such task is in assessing patients’ risk in undergoing a surgical operation by analyzing text that describes their history in other surgeries. Currently, this risk assessment is done with the assistance of human judges, who review unstructured surgical operation notes and other documents in order to extract relevant information.

Figure 1: STS Risk Calculator

The Society of Thoracic Surgeons (STS), for instance, has developed an online tool that, given enough information about patient demographics and reactions to previous cardiothoracic surgeries, assesses the risk of complica-
tions arising should the patient undergo another operation.

This project sought to analyze the sentiment of surgical operation notes based on factors that are more relevant to assessing future patient risk than the "good to bad" polarity useful for other domains such as reviews. Two factors were selected:

- How complicated a surgery was
- How concerned the surgical team may be assumed to be about the patient’s well-being after completing the surgery

These two factors are useful to assessing the future risk of a patient undergoing surgery, and were considered sufficiently distinct from each other to be evaluated separately.

2 Problem Description

Developing a new corpus is necessary because widely used sentiment lexicons and corpora are have only a limited effect in determining sentiment when applied to surgical operation notes. This is for two reasons:

1. Surgical operation notes are written to convey objective details of the surgery’s progress and thus, unlike domains for which large “good to bad” sentiment datasets exist such as movie reviews, are sparse in sentiment terms such as emotion words [1].

2. Many words in the medical domain, and more specifically in the domain of surgical operation notes, connote a different sentiment to what they connote in everyday domains. Thus, large sentiment lexicons such as SentiWordNet and the Subjectivity Lexicon can be used with only limited success for medical texts [1].

The two sentences below, parsed and annotated for "good to bad" sentiment by the Stanford Sentiment Analysis tool, illustrate the above limitations.

Figure 2: A sentence incorrectly labeled “good”

The sentence in Figure 2 describes the initial condition of the patient before surgery begins. Though doctors would not rate this condition as complicated nor concerning, they certainly would not label it “good” either. Yet the Stanford tool has done so, labeling the word “right” in “right coronary artery” as positive. Where word sense disambiguation is not used in sentiment analysis, it might be reasonable for the Stanford tool to default to labeling “right” as positive, but in the domain of surgical operation notes, “right” is used far more often for its sense as a direction and thus this default labeling is inappropriate.

Figure 3: A sentence correctly labeled “good”

The sentence in Figure 3 describes a successful step in the surgery, but the tool labels as negative several words that are used in this sentence in a neutral and objective manner to describe the procedure, such as "cardiac" and "cold blood." A domain-specific lexicon for surgeries would avoid these problems by having words scored
3 Approach

3.1 Selection of Sentiment Categories

“Complicated” and “concerning” were selected as sentiment categories for surgical operation notes as a domain-specific analog to the “good to bad” polarity used in other domains. This decision was based on an intuition that doctors who describe a surgery as “bad” are indicating one or both of the following:

- That the surgery was more involved or difficult than usual
- That the patient was unstable or in some other worrying condition

The words “complicated” and “concerning” were chosen to encapsulate the meaning of each of these ideas, and the following four sentences illustrate how the two constitute distinct and independent categories by which to score sentences in a note:

Sentence A) describes a routine procedure that leaves the surgical team unconcerned for the patient’s safety. Sentence B) describes that a patient was found to have “moderate calcification” in the coronary artery, and while this entails enough deviation from routine procedure to be labeled “complicated” by a doctor, it does not leave the surgical team concerned for the patient’s stability during or after the procedure.

Sentences C) and D) are concerning. Sentence C) describes a patient being transferred to the Intensive Care Unit and thus indicates that the patient is unstable, but the surgical procedure itself does not become more involved, so a doctor would not rate this sentence as “complicated.” In contrast, sentence D) describes a the patient becoming unstable in a way that necessitates a drastic deviation from normal protocol, as the patient’s heart has ceased contracting.

3.2 Annotating Tool

In order to create an annotated corpus of surgical operation notes, as well as a lexicon of sentiment words for this domain, an online annotator applet was created with JavaScript and given to doctors with access to YNNH operation notes. Operation notes had been extracted...
and converted to JSON files with fields for identifying information such as name and age as well as a field called “text” for the unstructured note. These JSON files were preprocessed using Apache OpenNLP to determine sentence boundaries. The annotator was directed to select a preprocessed JSON file and input it into the annotator tool. Annotators were introduced to two binary polarities, not complicated or complicated and not concerning or concerning. Two possible values, 0 and 1, were permitted for each factor so as to maximize annotator agreement. The annotators were instructed to label sentiment by three levels as follows:

1. Label any individual words or sequences of words that indicate that the surgery was progressing in a way that was complicated or concerning. Assume that unlabeled words will be scored by default to be both not complicated and not concerning, but label any words or sequences of words that especially indicate that the surgery was progressing in way that was not complicated or not concerning.

2. Score each sentence in the note for both categories. Does the sentence indicate that the surgery was progressing in a way that was not complicated, or in a way that was complicated? In a way that was not concerning, or in a way that was concerning?

3. Score the whole note for both categories. Overall, does it indicate that the surgery was not complicated, or that it was complicated? That it was not concerning, or that it was concerning?

After completing a note, annotators were instructed to save their labels as a JSON file.

A total of 10 notes were annotated, which altogether contained 763 sentences and 1557 words.

### 3.3 Supervised Learning

Supervised learning of the “complicated” and “concerning” categories at the sentence level was executed using an RNN implemented with LSTM units in Keras, as well as a Naive Bayes Classifier as a baseline result. All code was written in Python using the Jupyter Notebook App.

Three datasets of sentences and labels were used for training models and classifiers:

1. Sentences from operation notes labeled “0“ for “not complicated” or “1“ for “complicated“

2. Sentences from operation notes labeled “0“ for “not concerning“ or “1“ for “concerning“
3. Sentences from the Cornell Movie Review Dataset labeled “0” for “positive” or “1” for “negative” [4]

The movie review dataset was chosen because it is a large—5331 positive and 5331 negative sentences—and widely used dataset for binary classification tasks in sentiment analysis. The dataset was preprocessed to give semantically negative sentences the same label as semantically negative sentences in surgeries—that is, those that were scored as “complicated” and “concerning.” This was done in order to evaluate how models trained on the Cornell set performed in evaluating surgical note sentiment as another baseline.

For Naive Bayes Classification, annotated sentences were divided into a training set of 90% of the data, and a test set of 10% of the data. A dictionary of words in the sentences and classifier were created using the Natural Language Toolkit (NLTK).

The LSTM network was created as a Sequential() model in Keras, and contained two dropout layers of 0.2 each to prevent overfitting. The network was optimized with the Adam Optimizer. Datasets were divided into a training set of 80% of the data and a test set of 20% of the data. Furthermore, the training set was divided such that 90% was used for training and 10% for validation, leading to a final distribution of 72% training, 8% validation, and 20% test.

For each unique scheme of training, validation, and test set, two different embedding layers were used in two experiments. In one, embeddings of 128 features were initialized to random values and trained on the words in the input sentences.

To improve performance, pre-trained word embeddings were imported for the second experiment. While vast collections of pre-trained word vectors such as Word2vec and GloVe are available to the public, it has been shown that word embeddings trained on text in the biomedical domain, such as Electronic Health Records and biomedical articles, perform better in NLP applications on medical text compared to word embeddings trained on text that is not domain-specific [2].

Word vectors were imported from BioNLP, whose collection of 5,443,656 word2vec-style word vectors with 200 features was trained on PubMed and PMC biomedical papers, as well as on a Wikipedia dump [3]. While the BioNLP collection contains vastly more word vectors than are needed for the domain of surgical operation notes—including biomolecular terms, for instance—it covers words in surgical notes that...
would be considered obscure by conventional word vector collections. In fact, BioNLP recognized and had an embedding for nearly every word in the operation notes dataset.

4 Data Sets Used

Surgical operation notes were drawn from YNNH through the Yale School of Medicines Center for Outcomes Research and Evaluation (CORE). They were annotated by doctors affiliated with CORE. Some training was performed on sentences reflecting positive and negative sentiment from film reviews in the Cornell Movie Review Dataset [4].

5 Evaluation Method

Binary classifier performance was tracked over time by following training and validation loss, as well as training and validation accuracy, after each epoch of training. These values were plotted using the matplotlib library in Python. Models were tested on their own test sets, and additionally on other datasets, using common metrics such as accuracy, precision, recall, and F1 scores. A confusion matrix was also produced for each experiment.

6 Results

Initial tests on the Cornell set showed that, as expected, conventional sentiment analysis models on a “good to bad” polarity are not effective in determining sentiment for surgical operation notes. While the model trained on the Cornell set had a 0.74 testing accuracy on its own test set, it guessed operation note sentences’ “complicated” scores with 0.35 testing accuracy, and “concerning” scores with 0.36 accuracy.

Figure 8: Loss and accuracy over time for training on the Cornell Movie set over 50 epochs

Initial tests on the operation notes dataset revealed a high class imbalance in the scores—the vast majority of sentences in the operation notes were found by doctors to be both “not complicated” and “not concerning,” or a score of “0.” While the network learned from the balanced Cornell movie set, the number of sentences scored “1” in the operation notes was too small, and the network began assigning a score of “0” to every sentence:

Figure 9: Loss and accuracy over 50 epochs for training on the notes with “complicated” labels

The Naive Bayes classifier performed similarly when trained on “complicated” and “concerning” labels, predicting with high accuracy due to a low number of sentences scored “1”:

Given this issue, two methods were used to prevent classifiers from simply learning to label all sentences as “0.”

First, class balancing was implemented on
the training set to ensure an equal number of “0” and “1” scores.

Sentences in the dataset were sorted by score. Half of the “1”-scored sentences were placed in the training set, and a number of “0”-scored sentences equal to this was also placed there. The remaining sentences were placed in the test set, and the training set was shuffled before training. As a consequence of this, the training set size was equal to the total number of “1”-scored sentences in the set, which—especially for “concerning” sentences at 22—was too low for useful learning. Although the training curve appears to have learned some information about the sentences, testing accuracy remained low at around 0.5.

Second, a transfer learning scheme was attempted. A model trained on the large Cornell dataset with word vectors initialized to their BioNLP embeddings was then further trained on the operation notes sentences for “complicated” and “concerning” labels. While, again, the dataset labels were very unbalanced, the validation rate is seen to fall slightly over several epochs, indicating the network had learned to guess “1” scores at times.

Interestingly, it was found that models trained on “complicated” labels for operation note sentences performed nearly as well when tested on “concerning” test sets:

This indicates an overlap of sentences rated “complicated” and “concerning” in the annotated dataset, but a larger corpus of annotated sentences is needed before concluding whether the two factors are indeed correlated in most operation notes.
7 Conclusion

The as of yet small size of the annotated operation notes corpus, especially compared to corpora such as the Cornell dataset, presented a major obstacle to effective sentiment analysis on the notes. A significant factor that compounds the difficulty of the corpus’s small size is that, since the notes cover a wide variety of surgeries that describe different patient conditions and procedures, there may not have been significant overlap in vocabulary for a machine learning tool to find a signal.

Growing the corpus to several thousand sentences, like the Cornell set, shows promise for improving performance. It would be important when developing this corpus to avoid class imbalance by maintaining an equal bipolar distribution of sentences for both the “complicated” and “concerning” categories, even if this means losing sentences’ relations to others in the note, as the Cornell corpus does. One way this can be achieved is by having doctors create an equal number of fictional sentences for the polarities in each category. This would be more efficient than having them annotate a large number of sentences that are neither “complicated” nor “concerning,” and the resulting corpus, being fictional, can be published and made available to the public for further research without violating HIPAA.

Another option for exploration is to choose one general factor and have doctors rate it with a granularity of three, such as “good, neutral, bad.” Low granularity of two was selected for this experiment to maximize judge agreement, but allowing for three choices gives annotators greater room to distinguish between outcomes that are simply “not complicated” and “not concerning,” and ones that are truly desirable.

Besides improving sentiment analysis at the sentence level, there is also considerable room for analysis of operation notes at two other levels collected by the annotating tool:

1. The word and phrase level. With a significantly large collection of words and phrases annotated for sentiment for the “complicated” and “concerning” categories, a domain-specific sentiment lexicon may be created. This lexicon may be then used by tools such as the Semantic Orientation CALculator (SOCAL), which relies on a dictionary of lemmatized sentiment terms scored by polarity and intensity and has been shown to work in a variety of domains [5]. In addition, SOCAL implements a negation search in scoring sentiment, and thus could correctly score a sentence such as “Hemodialysis proceeded without any difficulty” as “not complicated” [5].

2. The note level. Determining the “complicated” and “concerning” sentiment of an entire operation note as a function of sentence sentiment may not require a sophisticated machine learning approach. In the dataset created so far, whole notes rated “complicated” have a number of sentences rated “complicated” at various points in the note, while notes rated “not complicated” have little to no such sentences. With a larger corpus, some threshold of “complicated” sentences, either as an absolute number or proportion beyond which the whole note becomes “complicated,” may become clear. In contrast, notes containing “concerning” sentences are not rated overall as “concerning” unless there are “concerning” sentences at
the end of the note. These differences reflect intuitions on how “complicated” and “concerning” differ: a doctor remembers a surgery as having been complicated if enough complications occurred any time during the procedure, but remains concerned about the patient’s condition only if the surgery ended on a bad note.

While the results with a small dataset do not show significant performance, then, there is good reason to enlarge it. This can be achieved rapidly at CORE by building a fictional sentence dataset, or by recruiting a larger number of annotators and adding keyboard shortcuts and other improved ergonomics to the annotator tool. Given the dearth of annotated data in the healthcare domain, doing so would be well worth the effort.

References


A Data

The surgical operation notes from YNNH contain identifying information and cannot be disclosed under HIPAA.

B Code

The HTML, CSS, JavaScript, and Python files used for this project will be uploaded to the CPSC 490 on-line database by the end of reading period, as per the 490 requirements.

C Acknowledgements

My sincerest thanks go to Dr. Schulz for his guidance and support during this project, and to Professor Radev for his consultation. In addition, I would like to thank Patrick Young, Thomas Durant, and Nathan Price at the Center for Computational Health at CORE for welcoming me into their office and offering me further guidance this term.