Machine Learning Engineer Nanodegree

Capstone Proposal

Michael Fryar

September 30, 2018

Proposal

Domain Background

In 2016, 42,249 Americans died as a result of an opioid overdose[1]. To put that in perspective, that is more than the number of deaths in 2016 caused by firearms (38,658) or motor vehicle crashes (38,748)[2]. The epidemic has struck small towns and big cities; families struggling to make ends meet and families that seem like they're living the American dream. Certain parts of the country have been hit especially hard, including my native Kentucky.

The most effective known method for reducing overdose deaths is medication-assisted treatment (MAT) which has been shown in clinical studies to cut overdose deaths by 70 percent[3]. However, the persistence of stigmatized views of addiction as a moral failing rather than a medical condition among policymakers and the general public has fueled resistance to efforts to expand access to MAT and discouraged individuals struggling with addiction from seeking treatment[4].

Despite its importance, very little data is currently available on opioid-related stigma. This is in part because it is costly to measure stigma with traditional tools such as surveys, which may also underestimate the pervasiveness of stigma due to social desirability bias. However, the combination of new data sources, such as social media sites like Twitter, and new machine learning algorithms for analyzing unstructured text data has the potential to fill this gap.

Problem Statement

Using Twitter's Streaming API, I have built a dataset of more than 700,000 tweets that contain opioid-related keywords. Analyzing these tweets to determine whether they perpetuate a stigmatized view of opioid-use disorder represents a sentiment analysis problem, which is a subset of natural language processing problems. Although many sentiment analysis tools have been created in recent years, most of these tools focus on analyzing the positivity/negativity of text data. In this project, I will seek to develop a supervised method for coding the text of tweets for opioid-related stigma.

Datasets and Inputs

To build a dataset of unstructured text data that can provide insights on the pervasiveness of opioid-related stigma, I have been tracking conversations that mention opioids and related terms on Twitter. More than 65 million Americans use Twitter each month and, unlike Facebook posts, tweets are public by default. This provides a large, publicly available dataset that is less likely to suffer from social desirability bias than traditional surveys[5].

Specifically, I have used Twitter's Streaming API to build a sample of more than 700,000 tweets that include the following keywords: burprenorphine, carfentanil, codeine, fentanyl, heroin, injection site, hydrocodone, methadone, morphine, naloxone, naltrexone, narcan, narcotic, needle exchange, opana, opiate, opioid, opium, overdose, oxycodone, oxycontin, percocet, suboxone, safe injection, supervised injection, vicodin, and vivitrol. Building on the work of others who have used machine learning to code unstructured text data to determine if it perpetuates stigmatized views of Alzheimer's disease [6] and rape culture [7], I will be manually coding a sample of this data and using it to train a classifier that can be used to code the remaining data for opioid-related stigma.

Solution Statement

In order to train a classifier to code tweets for stigma, I will be using two different supervised learning techniques.

1. tf-idf + Logistic Regression, Naive Bayes, SVM, Random Forest, Adaboost
2. Word2Vec + CNN

The first technique will process the tweet data by converting the text to a matrix using a term-frequency times inverse document-frequency weighting (tf-idf). This weighting reduces the impact of the most frequently occurring terms which are less informative. The processed data will then be analyzed with a variety of classifiers and I will select for my final model the classifier with the best performance on the validation set.

The second technique will process the tweet data using a different transformation known as Word2Vec[8]. Rather than converting the text into a matrix, Word2Vec creates vector representations of words in a high-dimensional vector space. By outputting vectors, Word2Vec allows the processed data to be fed in to a convolutional neural network (CNN).

Benchmark Model

One of the challenges in creating a custom classifier is the lack of a benchmark model. Following the example of Oscar, Fox, Croucher, Wernick, Keune, & Hooker (2017), I will compare the results of my classifier with the predictions of a simple majority classifier (also known as the Zero Rule or ZeroR classifier). If a majority of tweets do not contain stigmatizing content, then a simple majority classifier would predict that all tweets do not contain stigmatizing content.

Evaluation Metrics

To quantify the performance of the benchmark model and the solution model, I will use three metrics: accuracy, F-1 score, and area under receiver operating characteristic (ROC) curve.

- **Accuracy formula**

```
(TP + TN)/(TP + TN + FP + FN)
```

- Accuracy measures the proportion of hand-coded values that are correctly predicted. In the formula above, TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

- **F-1 score formula**

```
(2 * precision * recall)/(precision + recall)
```

- F-1 score takes the harmonic mean of precision and recall.
  - Precision tells us what proportion of tweets we classified as stigmatizing actually were stigmatizing. It is defined as `TP/(TP + FP)`.
  - Recall tells us what proportion of messages that actually were stigmatizing were classified by us as stigmatizing. It is defined as `TP/(TP + FN)`.

- **Area under the ROC curve** is the area the curve which is created by plotting the true positive rate against the false positive rate at various threshold settings.
  - True positive rate is identical to recall, which is defined above.

  - False positive rate tells us what proportion of messages that weren't stigmatizing were classified by us as stigmatizing. It is defined as `FP/(FP + TN)`

Project Design

This first step in my project workflow will be cleaning the raw data from Twitter. Since this is messy, real world data, a number of processing steps in Pandas will be required.

- Ricky Kim has an excellent list of steps for cleaning tweets for sentiment analysis that includes removing HTML, @ mentions, URL links, UTF-8 BOM, hashtags and numbers[9]
- Remove tweets from users without an identifiable location (since I am ultimately interested in measuring how stigma varies across the country).
- Remove tweets that seem to be from bot accounts based on frequency

After data cleaning, I will select a random subset of the data to be the training set and create coding instrument in Google Forms. Following Baum, Cohen, & Zhukov (2018), I develop this coding instrument based on a review the literature on opioid-related stigma to determine the main categories of stigma and break them down into components. I will then ask friends to help me in coding this data, making sure that each coder receives a common set of articles so that intercoder reliability can be measured.

Once I have a labeled training set, I will process the data and train classifiers following the steps laid out in the **Solution Statement** section above. Then, I will use both k-fold cross-validation and randomly-repeated cross-validation to measure the out-of-sample predictive performance of the classifiers. Finally, I will use the classifier to automatically code the rest of data set.

References

1. Seth, P., Scholl, L., Rudd, R.A., Bacon, S. (2018). Overdose Deaths Involving Opioids, Cocaine, and Psychostimulants — United States, 2015–2016. MMWR Morb Mortal Wkly Rep 67, 349–358.
2. Xu, J., Murphy, S. L., Kochanek, K.D., Bastian, B., & Arias, E. (2018). Deaths: Final Data for 2016. National Vital Statistics Reports, 67(5). Hyattsville, MD: National Center for Health Statistics.
3. Sordo, L., Barrio, G., Bravo, M. J., Indave, B. I., Degenhardt, L., Wiessing, L., Ferri, M., & Pastor-Barriuso, R. (2017). Mortality risk during and after opioid substitution treatment: systematic review and meta-analysis of cohort studies. The BMJ, 357, j1550.
4. Olsen Y., & Sharfstein, J. M. (2014). Confronting the Stigma of Opioid Use Disorder and its Treatment. JAMA, 311(14), 1393–1394.
5. McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., & Spiro, E.S. (2015). Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing. Sociological Methods & Research, 46(3), 390 - 421.
6. Oscar, N., Fox, P. A., Croucher, R., Wernick, R., Keune, J., & Hooker, K. (2017). Machine Learning, Sentiment Analysis, and Tweets: An Examination of Alzheimer's Disease Stigma on Twitter. Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 72(5), 742-751.
7. Baum, M. A., Cohen, D. K., & Zhukov, Y. M. (2018). Does Rape Culture Predict Rape? Evidence from U.S. Newspapers, 2000-2013. Quarterly Journal of Political Science, 13(3), 263-289.
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. In Advances In Neural Information Processing Systems, 3111-3119.
9. Kim, Ricky. (2017, December 18). Another Twitter sentiment analysis with Python — Part 2 [Blog post]. Retrieved from https://towardsdatascience.com/another-twitter-sentiment-analysis-with-python-part-2-54913