Ryan Crowley, Alejandro Salinas

Artificial Intelligence: Principles and Techniques

CS 221 Final Report

13 December 2019

Triaging Concussions using Natural Language Processing

Despite immense improvements in our understanding of traumatic brain injuries, the diagnosis and treatment of concussions remains woefully inadequate. Each year, there are an estimated 1.6 to 3.8 million sports and recreation related concussions in the U.S. and on average 10% of all contact sport athletes sustain concussions every year (Daneshvar 2010). With the use of imaging tests — such as MRI scans — as well as baseline measurements like the ImPACT test, concussions have become more easily diagnosable. However, the access and affordability of such tests remain a barrier for underserved populations and concussions remain difficult diagnoses for doctors; hence, many concussions go untreated every year.

We attempted to address this gap in care by harnessing the tools we have learned in CS 221 to build a supervised learning model that uses emergency room patient data, collected from the National Electronic Injury Surveillance System (NEISS), to predict the presence or absence of a concussion. This model could in theory be implemented as part of a triage system in emergency rooms. That is, during the intake of a patient at an emergency room, a nurse or other medical assistant could obtain a brief description of the patient's case as well as other demographic data. With this information, the model would output whether it is likely that the individual has a concussion and the patient would be treated accordingly.

Our baseline is a logistic regression model that utilizes patient demographics including age, sex, and race as features but does not incorporate the patient narrative. Running the logistic

regression model on the data, we find that the accuracy of the model does significantly better than randomly guessing with a test error of 33%. For our oracle, we enlisted the help of a Stanford-trained physician. The physician labeled a subset of the data consisting of 40 patients and achieved a test error of 30%.

Data Set and Cleaning Methods

Each year, the NEISS releases a dataset based on a representative sample of hospitals visits in the U.S. For our project, we queried for NEISS data from 2017 and 2018 and restricted the data to only contain information on patients who visited emergency rooms with head/neck-related injuries. Stored in csv files, the data is read into a vector of tuples, where each tuple contains: (narrative, age, race, sex, label). If the patient was diagnosed with a concussion, the label is 1; otherwise, the label is either -1 or 0 depending on the loss function.

Since some narratives reveal the doctor's diagnosis or contain inconsistencies, our model cleans the data through a series of approaches. Most importantly, the model removes words that reveal the diagnosis, such as 'concussion', 'laceration', and 'closed head injury'. The model also removes all special characters and punctuation, and replaces certain abbreviations with the full word in an attempt to maintain consistency, for example 'rt' is replaced with 'right' and 'h/a' with 'headache'. Data is read in from two csv files — one of which serves as our training dataset and the other as our testing dataset. Each file contains data on at least 60,000 patients.

Model implementation

Our implementation follows a modified version of a linear predictive model as discussed in class and is implemented in Python without any advanced libraries. First, the model runs a feature extractor that splits each narrative string into its individual component words and uses a sparse vector to represent the feature set where each word is mapped to the number of times it

occurs in a particular narrative. The feature set also includes the race, sex and age of the patient, where age is implemented as a categorical variable. The model then uses stochastic gradient descent to train by iterating through the training data and minimizing the corresponding loss function. The step size and number of iterations differ based on the choice of loss function. Once completed, the algorithm returns a learned weight feature vector that is tested on both the training and testing datasets. Hence, the model relies on a unigram approach to learn how to predict if future narratives are concussions. We also implemented a feature extractor that takes into account two word features via a bigram approach. In addition to utilizing hinge loss and logistic loss functions, we devised an alternative loss function for our model. In a typical formulation of the logistic loss function, y takes on a value of 0 if no concussion is present and a value of 1 when a concussion is present and can be written as:

$$w_{i+1} \leftarrow w_i - \eta * \phi(x) * \frac{1}{1 + e^{(-\phi(x) \cdot w_i)y}}$$

Since we determined that it is more important that our model be able to identify concussions, at the expense of predicting concussions when they aren't present and since our dataset has many more cases of non-concussions than it does of concussions, we devised an altered loss function based on the logistic loss function. The loss function update is as follows, where $1[y == 1]$ takes on a value of 1 when y = 1 and a value of 0 when y = 0:

$$w_{i+1} \leftarrow w_i - (\alpha * 1[y == 1] + 1) * \eta * \phi(x) * \frac{1}{1 + e^{(-\phi(x) \cdot w_i)y}}$$

Here, α serves as a hyperparameter that indicates the extent to which the magnitudes of the directions of the gradient for training points of concussions should be prioritized over training points of non-concussions. In this case, $\alpha = 0$ is equivalent to logistic loss whereas increasing values of α relate to increasing preference for concussions which then leads to the model predicting concussion at higher rates.

Additionally, we constructed a neural network from scratch in Python. Since we implemented it without using advanced libraries, our deep neural network is necessarily simple, consisting of a single hidden layer and a sigmoid activation function. Additionally, this model maintains the same basic structure of the models described above, where it takes in patient demographics and narratives as features and outputs whether or not a patient is likely to have a concussion. Pre-trained word embeddings were not utilized due to the massive nature of the datasets present and the relative narrowness of this classification task.

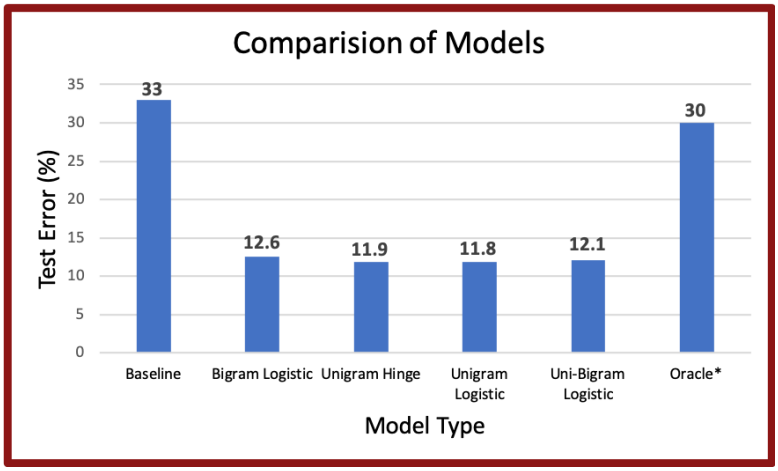
Primary Analysis

We compared the performance of different models using test error as the measure of optimal performance of the model:

$$Test\ Error\ = \frac{False\ Positive + False\ Negative}{True\ Negative + False\ Positive + False\ Negative + True\ Positive}$$

Models were trained on 2017 NEISS data (n=63,978) and tested on 2018 NEISS data (n=61,436).

Figure 1: Model Comparison



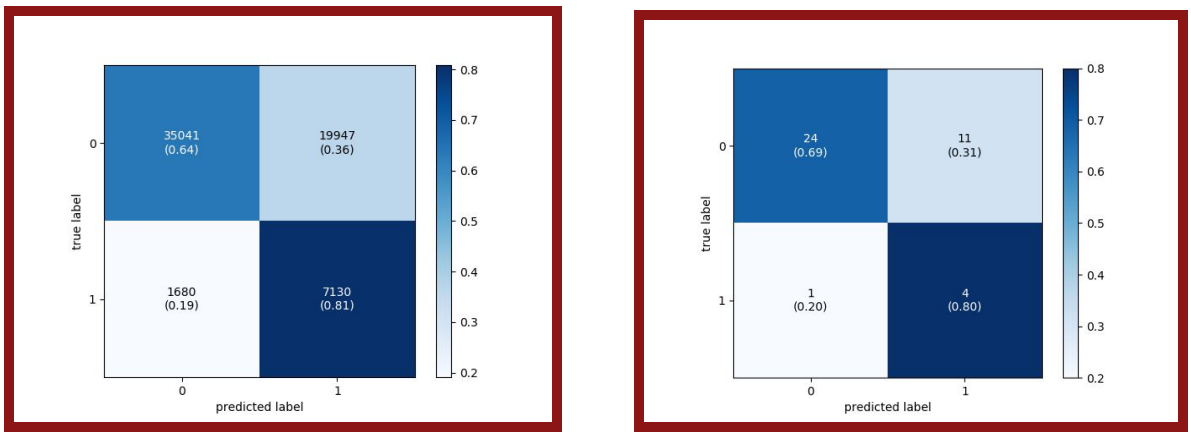
*Due to the infeasibility of getting the physician to label a large dataset, a separate dataset (n=40) was used to test the physician's performance.

From the results, we can see that the Unigram model with the logistic loss function

achieved the optimal test error of 11.8% with the other models also demonstrating similar performance. We can see that this marks a significant improvement over the baseline. To gain deeper intuition into the types of predictions that the model was making we also examined the confusion matrices associated with the performance of the models. This allows examination of the types of errors that the model makes. Some useful statistics for this analysis are as follows:

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \qquad Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

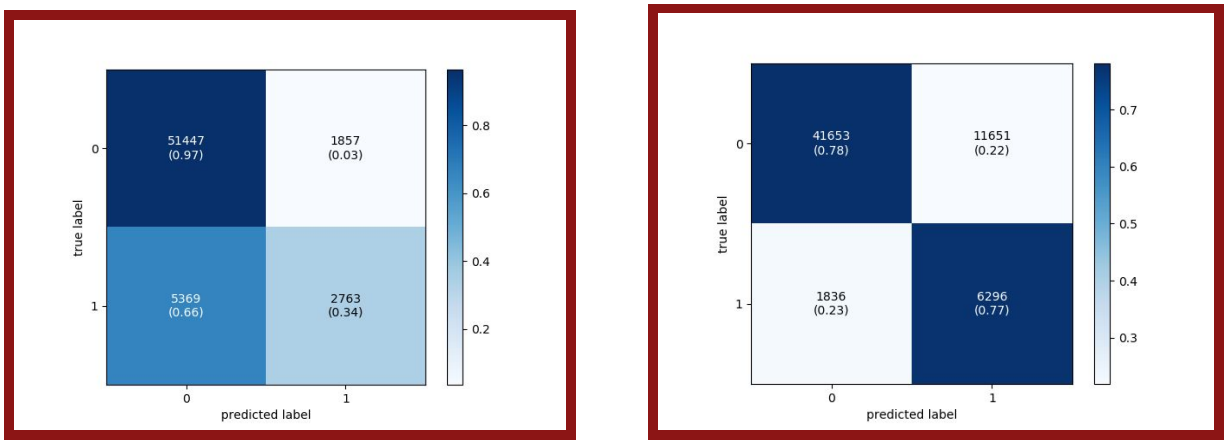
Figure 2: Confusion Matrices for Baseline (left) and Oracle (right)



Here, we can see the baseline tends to have a greater difficulty of identifying patients that don't have a concussion and hence suffers from lower specificity. Overall, the baselines and oracle are fairly comparable in their performance indicating the difficulty of this task for humans.

It's important to note the dataset is not split evenly between concussions and no concussions. Hence, it's possible for a model to achieve low test error by always predicting the latter. But, this is a useless model as no learning is needed to predict the same outcome every time. While our models did not always predict no concussion, some did predict no concussion significantly more often; hence, these models were not effective at detecting concussions. Accordingly, we also employed our altered loss function for prediction and assessed the results.

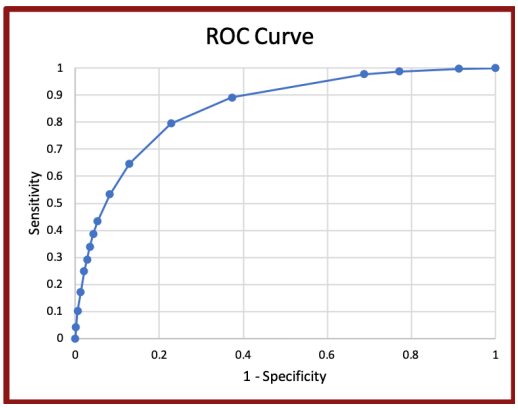
Figure 3: Confusion Matrices for Unigram Logistic Loss (left) and Altered Loss Function (right)



We can see that both models perform well and outperform both the oracle and the baseline model in relation to model error. However, the model on the right — with a loss function that prioritizes gradients of training examples where the outcome is a concussion — achieves a greater balance of sensitivity and specificity as .77 and .78, respectively, whereas the logistic loss function only achieves a sensitivity of .34. Hence, we conjecture that the altered loss function leads to a better classifier for triaging concussions. Note that for the altered loss function, we set $\alpha = 1.5$ as chosen by a validation set approach of training on the 2015 NEISS data and testing on the 2016 NEISS data.

To continue exploring the tradeoff between sensitivity and specificity, we trained and tested the unigram logistic model on a variety of different thresholds, and obtained a Receiver Operating Characteristic (ROC).

Figure 4: ROC Curve for Logistic Loss Unigram



This curve describes the intriguing tradeoff between sensitivity and specificity for the logistic loss unigram model. Varying the threshold is important in this context as the two possible errors that the model

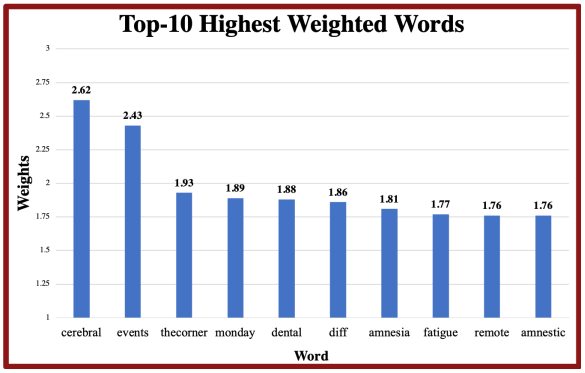
can make are not equivalent as missing a possible concussion this early in the triage process is a much more grave error than falsely predicting a concussion when none is present. We conjecture that a model of this form employed in the emergency room setting should seek to optimize sensitivity, at the expense of specificity, ensuring a smaller proportion of concussions are missed.

Additionally, the neural network that we trained was able to achieve a test error of .325 on a training dataset (n = 300) and a test dataset (n = 119) much smaller than the datasets used to examine the other models but performed extremely poorly on the larger datasets.

Exploratory Analyses

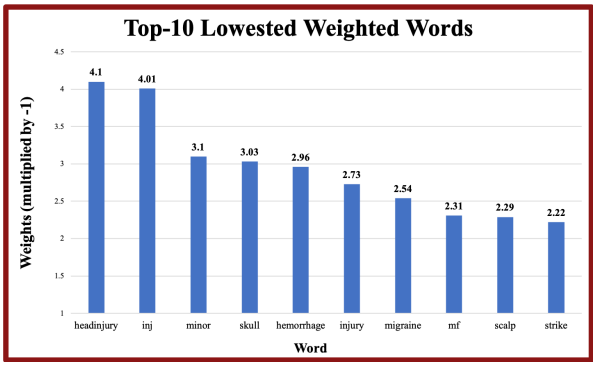
After training a logistic loss unigram model on NEISS data from 2017, we tested the model on 2018 data. The figures below reflect explorations into interpretations of our results.

Figure 5: Top-10 Highest Weighted Words



This figure lists the 10 words with the most positive weights, representing the collection of words most positively associated with a concussion based on the data. Words such as 'amnestic', 'amnesia' and 'fatigue' are particularly interesting because their relatively high weights suggests a correlation with concussions. Other words such as 'thecorner', 'diff', and 'monday' may suggest there is noise in our dataset.

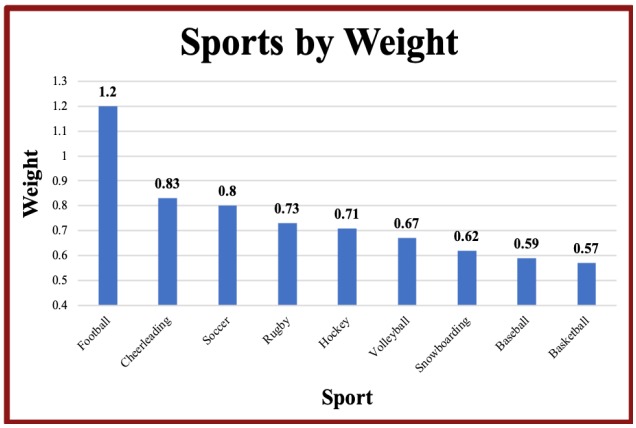
Figure 6: Top-10 Lowest Weighted Words



This figure lists the 10 words with the most negative weights, representing the collection of words most negatively associated with a concussion based on the data. Words like 'minor', 'headinjury',

'migraine' and 'hemorrhage' all suggest patients have injuries and pain not related to concussions but still severe enough to seek medical attention at the emergency room.

Figure 7: Sports by Weight



Depicted here are the weights of key sports. Football is most positively associated with a concussion, and its difference in weight to other sports is quite revealing of the potential danger of the sport. A graph like this could help argue for safer protocols in these sports to lower the number of incidences of concussions.

Social Impact

Our results demonstrate that a model taking into account patient demographics and a brief narrative of the events leading up to the accident can create an effective tool for triaging concussions. Intriguingly, we found that human performance was bested by the models that we developed. This is not to say that a model of the type proposed should replace human diagnosis for concussions; in practice, physicians have access to a much broader set of signs and symptoms when conducting a physical examination and hence will be able to achieve much lower errors in predicting concussions. Rather, the superior performance of our model to a physician indicates both the difficulty of this triage task with limited data as well as the potential usefulness of the model proposed in triaging patients to assess their risk of concussion.

The importance of the interpretability of the developed logistic and hinge loss models can't be underestimated. Because the model consists of a single weight for each individual feature, the model allows for the uncovering of statistical associations between words and the presence of concussions that may have not been previously known. For instance, we found it

fascinating that the words ‘fatigue’ and ‘amnestic’ were very positively associated with concussions, indicating that these two symptoms are likely indicators of the presence of a concussion. Furthermore, an interpretable model avoids the possible philosophical quandaries associated with a ‘black box’ model. That is, some argue that an uninterpretable “black box” model in medicine actually contravenes the moral obligations of physicians (London 2019). Hence, models with higher degrees of interpretability, such as logistic regression models or decison trees tend to be favored in medicine.

We were disappointed to see that the neural network we developed was unable to achieve comparable results to the performance of the other models on training datasets consisting of thousands of patients. It is likely that a structure with more complexity is necessary for the neural network to perform optimally on this task. With more time, we would continue to finetune this neural network. In particular, recurrent neural networks have been pointed out to be an especially common and effective tool in clinical natural language processing (Wu 2019).

Overall, we found that the models developed were effective in their ability to accurately triage concussions. This demonstrates the potential practicability of this approach for implementation in real world settings. Furthermore, we recognize that a model of this form can be used for triaging patients with other diseases and injuries as well. Considering the difficulty that remains in effectively diagnosing and treating individuals with concussions, it is imperative that we work towards finding better solutions. A model that can help physicians recognize patients who are more likely to have concussions could in theory help decrease the number of concussions missed and improve a hospital’s overall care of such individuals. Beyond the data, we should strive to keep in mind that each concussion missed is not a statistic but rather a monumentally devastating mistake that has a profound impact on the individual affected.

Works Cited

Daneshvar, Daniel H., et al. “The Epidemiology of Sport-Related Concussion.” Clinics in Sports Medicine, vol. 30, no. 1, 2011, pp. 1–17., doi:10.1016/j.csm.2010.08.006.

London, Alex John. “Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability.” *Hastings Center Report*, vol. 49, no. 1, 2019, pp. 15–21., doi:10.1002/hast.973.

Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, Bo Zhao, Hua Xu, Deep learning in clinical natural language processing: a methodical review, Journal of the American Medical Informatics Association, , ocz200, https://doi.org/10.1093/jamia/ocz200

Link to Google Drive with Code.zip and Data.zip:

https://drive.google.com/drive/folders/1GaqQJ2z8bLgORj38bCc1BXVs2Fjm9gJO?usp=sharing