Triaging ER Concussions using Natural Language Processing

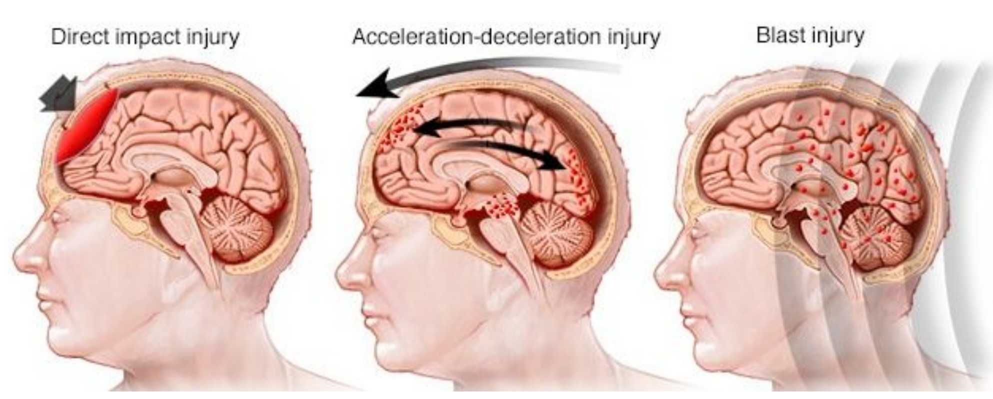
Alejandro Salinas and Ryan Crowley

CS221 Artificial Intelligence, Department of Computer Science, Stanford University

Stanford
Computer Science

Introduction

Each year, an estimated 1.6 to 3.8 million sports and recreation related concussions in the U.S occur. The access and affordability of treatment for concussions remain a barrier for certain populations and diagnosis remains difficult; hence, many concussions go untreated every year. Using a decade's worth of emergency room patient data that includes the patient's age, race, sex, and a brief narrative of the injury, we have constructed a supervised learning model that predicts the presence of a concussion.



Approach

Our baseline is a linear model that takes into account patient age, race, and sex. Our oracle is a Stanford-trained MD with access to each patient's demographic data and narrative. After cleaning the data, the models created use a unigram approach that breaks the narrative down into its constituent words which serve as features. Simpler linear models with hinge and logistic loss functions were fit using SGD. A more complex neural network was also developed. All code was developed in python without any advanced libraries.

Results

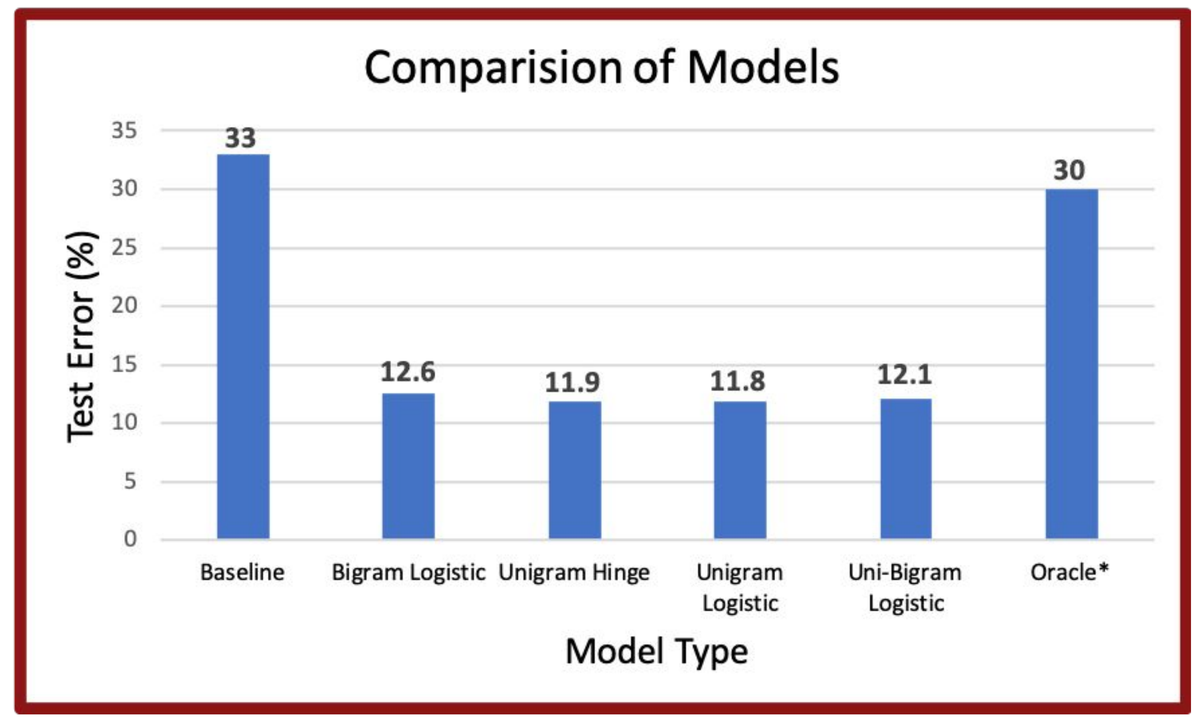


Figure 1. *This figure compares the performance of different models classifying concussions on the same train (n=63,978) and test (n=61,436) datasts.*
**Due to the infeasibility of getting the physician to label a large dataset, a separate dataset (n=40) was used to test the physician's performance.*

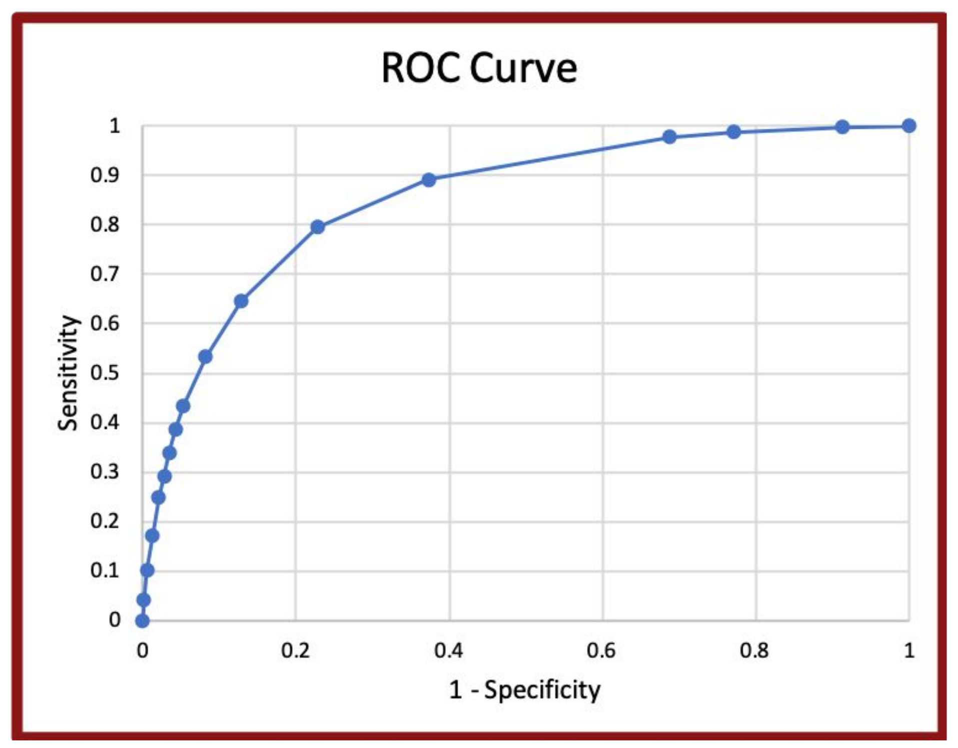


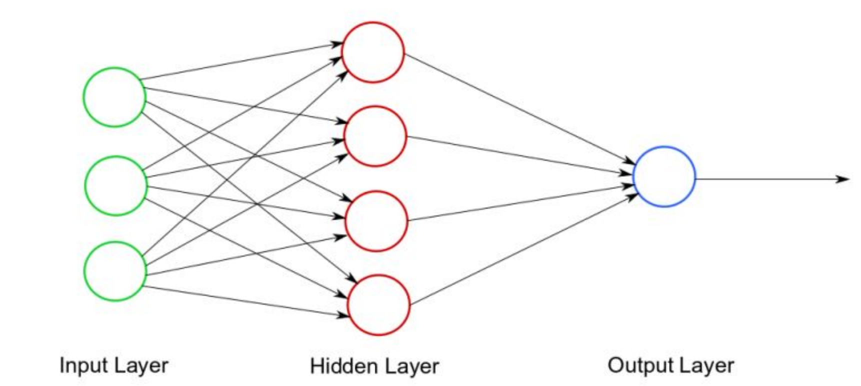
Figure 2. *This figure shows the Receiver Operating Characteristic Curve for the unigram model with a logistic loss function. This graph illustrates the tradeoff between sensitivity and specificity of the model and can be used to choose an optimal threshold for the model.*

TOP TEN	BOTTOM TEN
1. cerebral	1. headinjury
2. events	2. inj
3. thecorner	3. minor
4. monday	4. skull
5. dental	5. hemorrhage
6. diff	6. injury
7. amnesia	7. migraine
8. fatigue	8. mf
9. remote	9. scalp
10. amnestic	10. strike

Figure 3. *This figure lists the 10 words with the most positive and most negative weights of a unigram model with logistic loss function. The words in the top 10 category represent the collection of words most highly associated with a concussion; likewise, the words in the bottom 10 category are the collection of words most negatively associated with a concussion, based on the data.*

Artificial Neural Network

We implemented a simple artificial neural network with a sigmoid activation function and a single hidden layer in python. The model converged on smaller datasets but didn't converge when larger datasets were used. It obtained a test error of .325 with train (n=300) and test (n=119) datasets.



Discussion

We are impressed with the performance of the models. The unigram model demonstrates its usefulness as a triage tool. Intriguingly, the developed models outperform the oracle, indicating the difficulty of this prediction task without further diagnostic data. We also found that the model with both unigram and bigram features performs worse indicating that too many features adds unwanted noise to the data.

Future Steps

With more time we would continue to troubleshoot the neural network to find settings where it converges for the larger data sets. Additionally, we would work to include pre-trained word embeddings like Word2vec for this classification task.