Using natural language processing and machine learning to predict severe injuries classification in the oil and gas industry

Marcelo Guarido and Daniel Trad

ABSTRACT

Severe injuries, such as fractured body parts and amputations, are always on the top list of mitigation importance in any kind of job. In this work, we use the incident/accident description of the severe injuries' reports from the Occupational Safety and Health Administration of the United States Department of Labor to create a machine learning model that standardizes the class of the incident classification. We used natural language processing to convert each description of an injury to numerical features and applied the TF-IDF methodology to remove words that are not important to the classification of an injury. Models such as "Extremely Randomize Trees" and "Multinomial Logistic Regression" were trained and applied on the oil & gas industry's reports to test their accuracy, and we came to the following conclusions: predictions are improved when binary input features are used; the Extremely Randomized Trees tends to predict the most frequent classes with accuracy over 80%; the Logistic Regression works better for the other classes with balanced accuracy of 54% if implemented with balanced class weights.

INTRODUCTION

Natural Language Processing (NLP) is the use of text and/or speech data to analyze and predict specific goals of the research. We can find examples in our daily life, as for speech recognition in our phones (Weber 2002).

For health and safety, NLP has been used with impressive results during the last years. Dublin et al. (2013) use NLP to identify pneumonia from radiology reports, where it could classify with high accuracy 75% of the reports (the remaining 25% required manual review). Yetisgen-Yildiz, Bejan, and Wurfel (2013) use free-text chest X-ray reports to extract unigram, bigram, and trigram features to help in the prediction of acute lung injury, with precision higher than 80%. More on the injuries prevention side, Tixier et al. (2016) use injury reports from construction sites to extract valuable information and insights from a poorly structured data set and with difficult manual analysis capabilities. Chokor et al. (2016) propose the use of NLP and unsupervised learning on the construction sites injuries from the Arizona *Occupational Safety and Health Administration of the United States Department of Labor* (OSHA) to divide the injuries sources into main clusters (as fall and electrocution) to help in the prevention of incidents and in the improvement of the safety regulations.

In this work, our proposal is to use the *severe injury reports* from the Occupational Safety and Health Administration of the United States Department of Labor to create a NLP classification system that can standardize the injury classification for the oil & gas industry, as each company, region, and employee can write and classify the incident based on local language standards. We start the project doing the analysis of the data from all the industries and discriminating the oil & gas industry reports, and than show how to convert

an incident description into numerical features, ending up with the classification of the injury.

DATA DESCRIPTION

According to the Occupational Safety and Health Administration of the United States Department of Labor, the data contains severe injury reports information from January 1^{st}, 2015, to February 28^{th}, 2019. It is a `CSV` file with the following information about each incident (Table 1):

Table 1: Data description.

Name	Description
Identification Number	Incident identification number
UPA	Local agency number
EventDate	Date of the incident
Employer	Employer name
Address1	Address
Address2	Address
City	City of the incident
State	State of the incident
Latitude	Latitude of the incident location
Longitude	Longitude of the incident location
Primary NAICS	NAICS industry identification number
Hospitalized	0 means no and 1 means yes
Amputation	0 means no and 1 means yes
Inspection	Inspection number
Final Narrative	Incident description
Nature	Code of the type of injury
NatureTitle	Type of injury
Part of Body	Code for the part of the body injured
Event	Code for the event
EventTitle	Title of the incident
Source	Code for the source of the incident
SourceTitle	Source of the incident
Secondary Source	Code for the secondary source of the incident
Secondary Source Title	Secondary source of the incident

The data is fairly organized and well standardized, however some cleaning and fixing was necessary for addresses and coordinates. Apparently, some of the incidents had city, State, and coordinates mixed up. To fix that, we assumed that the *zip code* of each incident is correct, and we force all other address information to match the zip code.

With the data fixed, Figure 1 presents the location of severe injuries from 2015 for all the industries contained in the data (in orange), summing a total of 41541 injury reports, and for the companies classified as Oil&Gas (in red), with a total of 1351 reports. Most of the injuries are concentrated on the East side of the United States, as this is the most populated area of the country. The injuries related to the Oil&Gas industry seem to have less dispersion when compared to all industries, and are located in bigger centers for the industry. Another observation is that some states at the East coast have less injuries reports than the surrounding states. This could be the reality of the incidents, or it could be different regulations for report submissions.

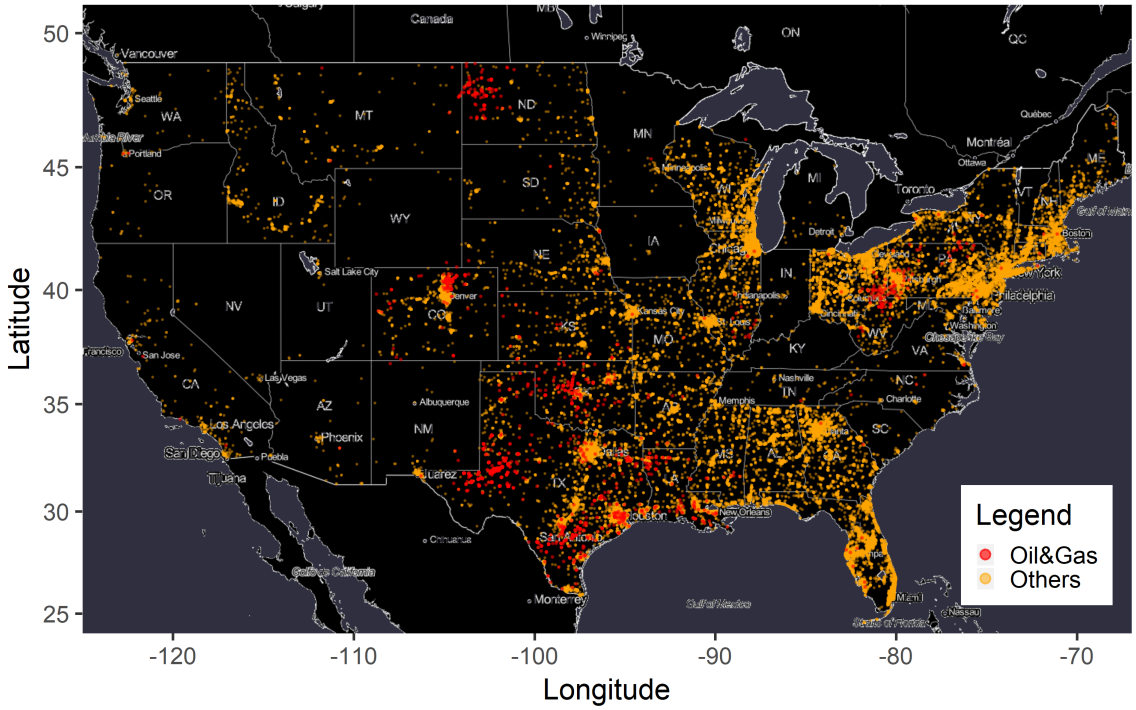

Figure 1: Map of all the severe injuries from 1/1/2015 to 2/28/2019, for all industries (orange) and for the Oil and Gas industry (red).

Figure 1 can give us a brief idea of the injuries around the country. But we can extract even more insights of the injuries behavior by simply making different plots.

10 most dangerous cities to work in

One of the analysis (and a straight forward one), would be to identify the 10 most dangerous cities to work in (in absolute numbers), divided by *Oil&Gas* and *all* industries, and by *hospitalization* and *amputations*.

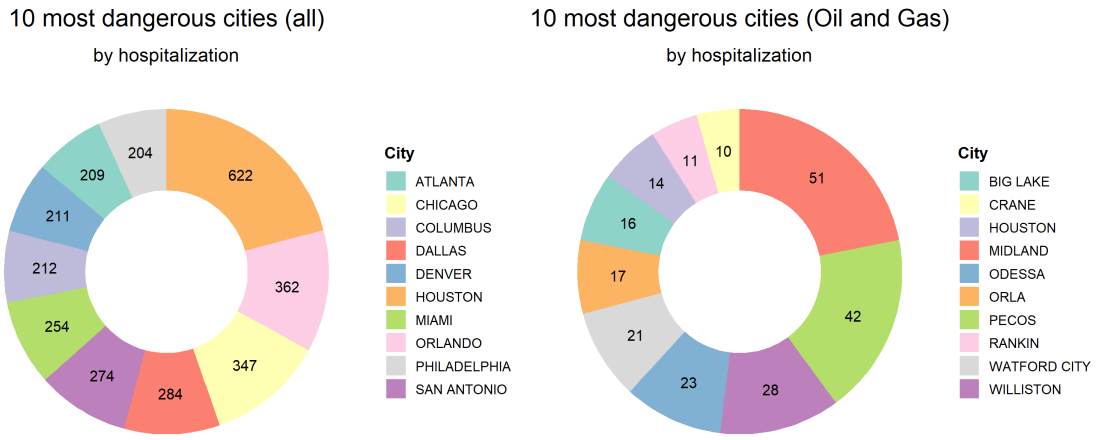

Figure 2: Number of reports of the 10 most dangerous cities (by hospitalization) for all the industries (left) and for the Oil and Gas industry (right).

Figure 2 shows an interesting insight: when looking at all industries combined, Houston is the most dangerous one (622 injuries, by hospitalization), but it goes down to the 8^{th} for the Oil&Gas industry. Actually, the most dangerous city for the Oil&Gas industry is Midland (51 injuries). That is curious as Houston is the the largest Oil&Gas center in the US, so anyone could expect Houston to be the most dangerous city for the industry. The explanation could be that Houston is more an office city. As this is a **severe injuries** dataset, the accidents are more related to the use of heavy machinery, which are usually located well away from those larger centers, or on production sites.

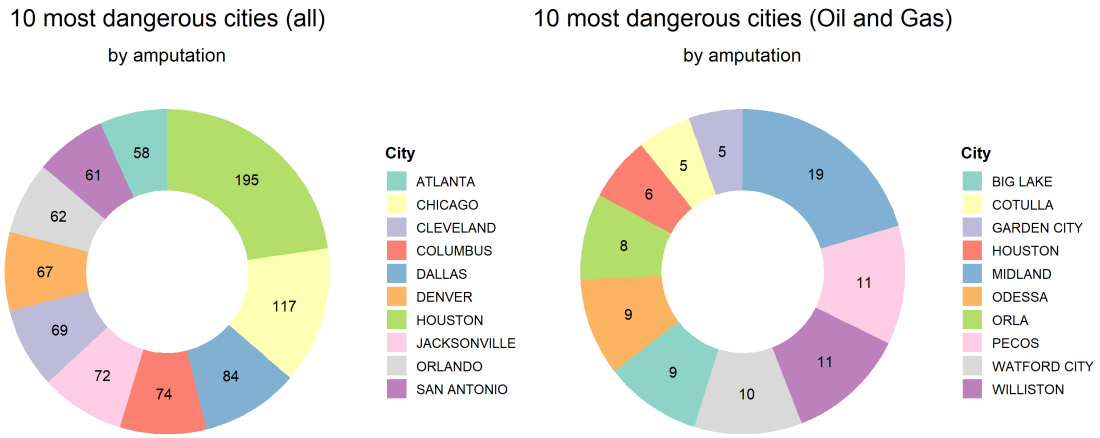

Figure 3: Number of reports of the 10 most dangerous cities (by amputation) for all the industries (left) and for the Oil and Gas industry (right).

When looking into *amputations* in Figure 3, for both plots, the behavior is similar to hospitalization, with small changes on the order of the top ranked cities, with some cities entering the top 10 rank (like Atlanta), and some leaving (like Philadelphia).

10 most dangerous states to work

After checking the most dangerous cities for workers, it is time to analyse the most dangerous states for workers. We expect that the behavior should be similar to the cities.

In Figure 4 is shown the 10 most dangerous states for workers when considering only hospitalization, for all the industries (left) and the Oil$Gas industry (right). In both scenarios, *Texas* is the most dangerous state. For the Oil&Gas industry, Texas has more hospitalization than the other 9 states combined. This analysis is not surprising, as most of the oil and gas exploration and production in the US is in Texas. Most of the companies are located in this Southern state, hence, most of the accidents will be there.

Figure 5 presents the 10 most dangerous states for workers when considering amputations. The analysis is similar to the one for Figure 4, with Texas at the top of the list for both situations (all and Oil&Gas industries). Only for all industries we see Ohio going to 2^{nd} place. But in general there are just smaller fluctuations in positions in both plots.

As a last observation, in both cases (hospitalization and amputation), Texas is still the most dangerous state of the US if we remove the Oil&Gas industry. Probably some safety

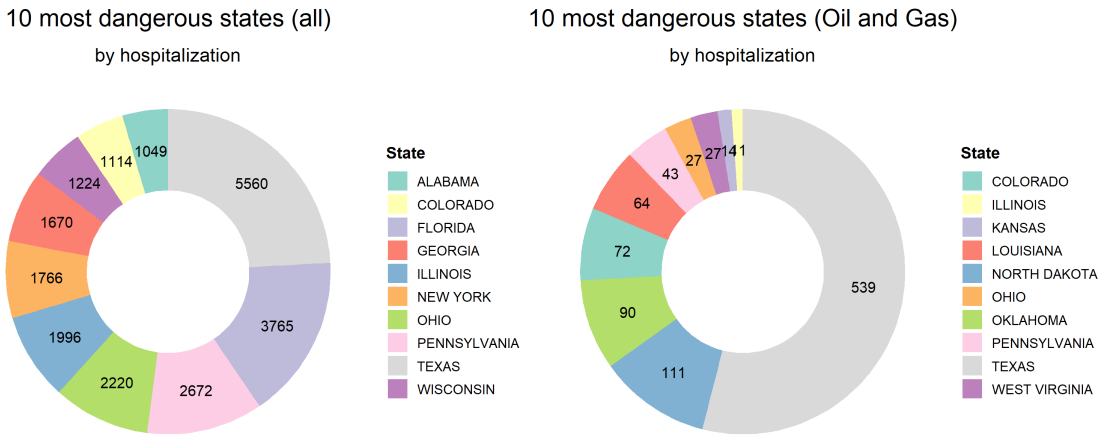

Figure 4: Number of reports of the 10 most dangerous states (by hospitalization) for all the industries (left) and for the Oil and Gas industry (right).

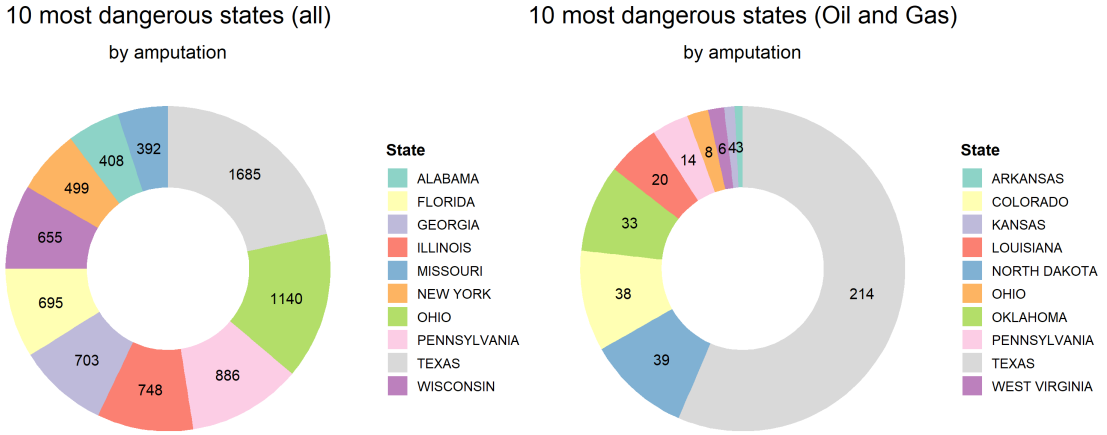

Figure 5: Number of reports of the 10 most dangerous states (by amputation) for all the industries (left) and for the Oil and Gas industry (right).

measures will need to be taken in the state to improve those indicators.

10 most dangerous companies to work

Now, let's take a look on the 10 most dangerous companies (employers) to work for in US. Again, the analysis is done separated for *all industries* and for the *Oil&Gas industry*, for both *hospitalization* and *amputations*.

First for hospitalization, Figure 6 shows the top 10 most dangerous companies to work for all industries (left), and for the Oil&Gas industry (right). For all the industries, there is a tendency of mailing employers (USPS in 1^{st} place and UPS in 3^{rd} place) to have more hospitalizations than any other employer (exception for Walmart, which is 2^{nd} place). Mailing employers are generally large companies, as well as Walmart, and work with storage, delivery, commute, and heavy machinery. It makes sense for them to be on the top of the list. And let's remember that those are *absolute numbers*, and not a proportion by the size of the employer. Bigger companies will tend to have higher numbers of injuries. There is, of

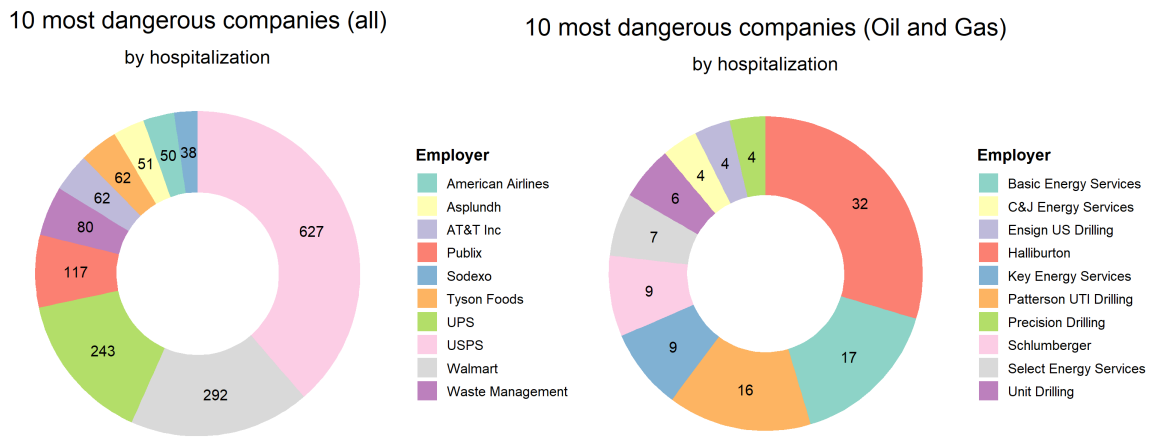

Figure 6: Number of reports of the 10 most dangerous companies (by hospitalization) for all the industries (left) and for the Oil and Gas industry (right).

course, also a correlation to the kind of work done.

Still in Figure 6, looking at hospitalizations, the top 10 employers are led by companies with drilling services, such as Halliburton, Basic Energy Services, and Patterson UTI Drilling. Halliburton is one of the largest offshore drilling companies, as Basic Energy Services and Patterson UTI Drilling are focused on onshore drilling. For Halliburton, we are confident to say that most of the injuries happen on the offshore rigs, at the Gulf of Mexico. The map in Figure 1 probably shows injuries located in the headquarters of the companies, so most of the injuries must have moved in-land, in Texas. Drilling services involve high peril, and the top 10 list points to this.

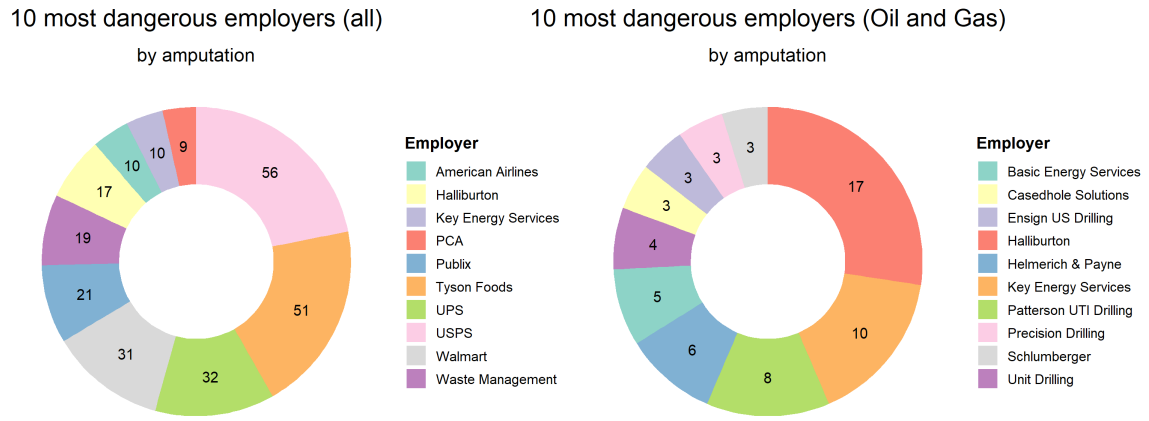

Figure 7: Number of reports of the 10 most dangerous companies (by amputation) for all the industries (left) and for the Oil and Gas industry (right).

Now, looking at amputations in Figure 7, for all the industries (left), the mailing employers are on the same position as for hospitalizations, but Walmart dropped to 4^{th} as Tyson Foods got the 2^{nd} place. This may be related to the type of work done at the employers location. Walmart has machinery to move heavy products, and an accident in this facility may not necessarily lead to amputation. However, dealing with food, there may have a larger number of cutting machines, and the resulting type of accidents can lead to a

larger number of amputations.

For the Oil&Gas industry in Figure 5 (right), drilling companies are still on the top of the list, with some fluctuation in the positions. But Halliburton continues as the top one.

Top 10 injuries

So far, the most dangerous cities, states, and employers were analyzed by counting the number of hospitalizations and amputations (specified columns of the dataset) for each group. Next follows an analysis of the top 10 types of injuries for all industries and Oil&Gas industry.

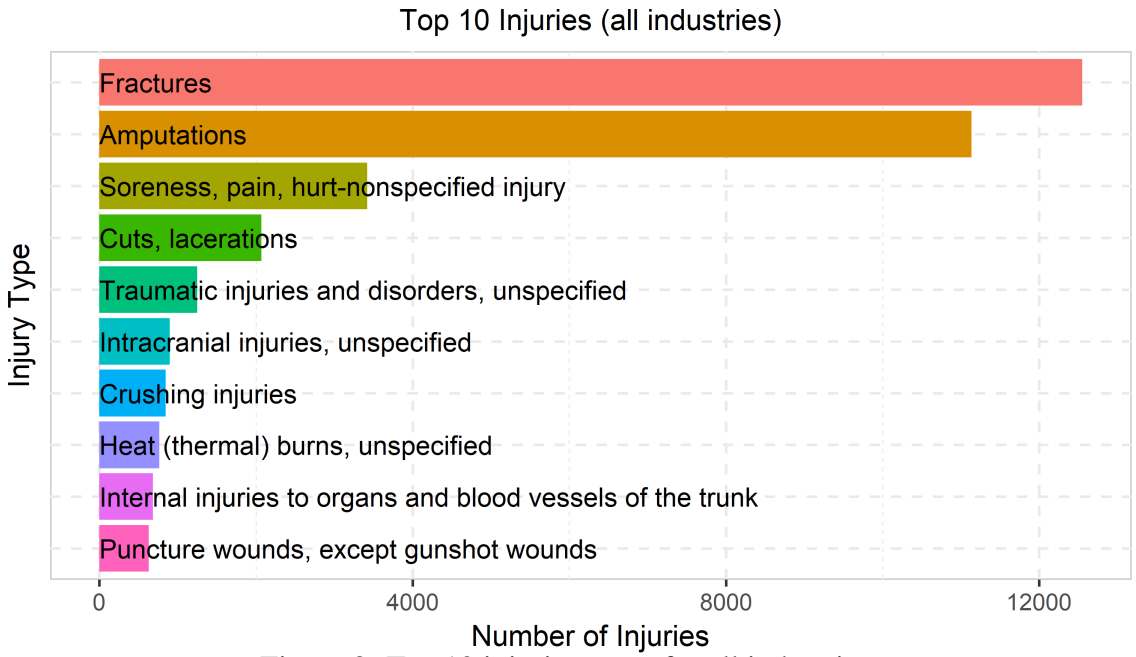

Figure 8: Top 10 injuries types for all industries.

Figures 8 and 9 show, respectively, the top 10 injuries for all industries and for the Oil&Gas industry. In both scenarios, the top 2 injuries are *fractures* and *amputations*, however, in different order for each case. For all the industries, fractures are the top injury type, and for the Oil&Gas industry, amputation looks to be more common (the difference is not large). Remember that for the top 10 employers for all industries, the main industry for injuries is the mailing one. The type of accident more common can be the ones related to traffic and impact accidents, leading fractures to the top of the list. For the Oil&Gas industry, the main employers in the list are the ones for well drilling, which machines have lots of moving parts and an accident can cause amputations.

Looking from the 3^{rd} to the 10^{th} positions in Figures 8 and 9, there are lots of similarities. The differences are the *intracranial injuries* that appear in a relatively high position (6^{th}) for all the industries, but is missing in the Oil&Gas list. We believe the use of *PPEs* (Personal Protective Equipment), such as helmets, may take a more important position in oil&gas safety regulation than in some others industries. For Oil&Gas, the most notable injury in the top 10 is the *poisoning, toxic, noxious, or allergenic effect* one. That makes

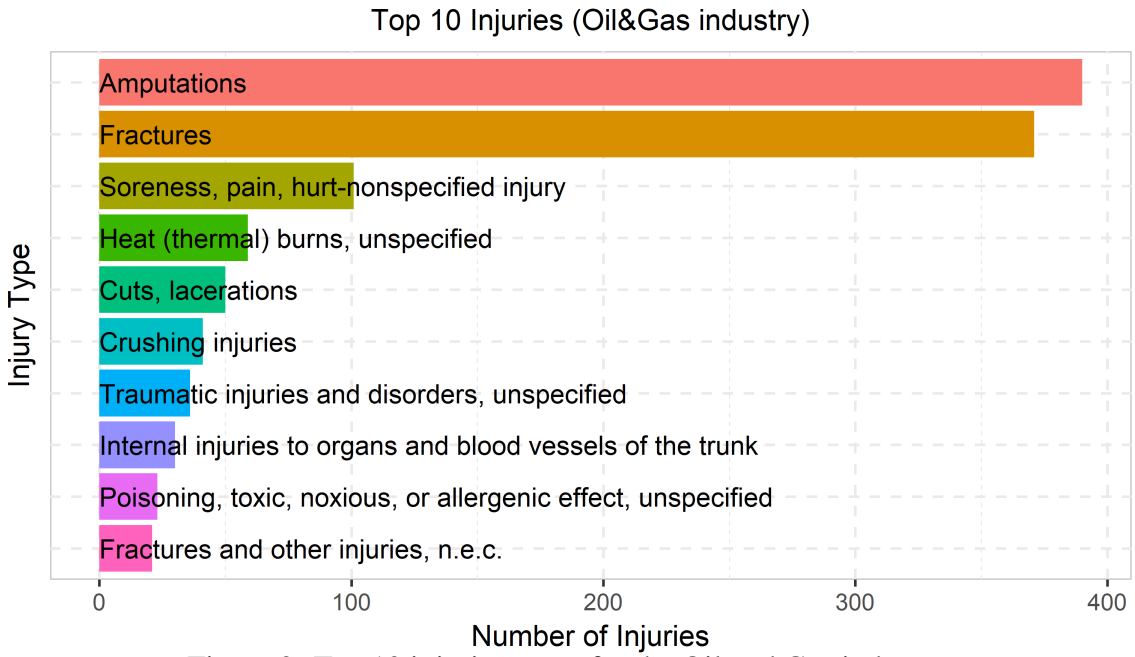

Figure 9: Top 10 injuries types for the Oil and Gas industry.

sense, as Oil&Gas companies involved in drilling and refining can be in contact with toxic materials.

NATURAL LANGUAGE PROCESSING

One question we can ask to the severe injuries dataset is: *"Can the dataset be used to create a ML model to standardize the injure type classification?"*. In other words, can we use the description of the incident to predict in which injury type class it belongs? Well, first, let's understand the importance of this question. Having a standardized dataset is helpful for data analysis, where all the incidents are classified with the same class name/number. Non-standardized data come from each company (or group of companies) having its own classification for an injury type. For example, let's say there is an incident classified as *fractured arm*. Some companies can classify it as *fractured arm*, others can classify it as a *broken arm*. Both have the same meaning, but from the data analysis point of view, the algorithm can interpret them as different classes, if no pre-processing is done. Looking at data sets with a large number of classes, the pre-processing becomes harder and misinterpretations are likely to happen.

One way simple but powerful way to do this classification is to use an approach similar to *sentimental analysis* (Shahana and Omman 2015). The idea is to count the words used in each incident description (converting the string to a numerical variable, or feature), and then giving weights to each word. However, some words may appear in every or most descriptions, such as *the*, *an*, *he/she*, etc. Luckily, there is a method that helps eliminating these common words in the analysis, only keeping *important* words. The method is called *term frequency–inverse document frequency*, or simply *TF-IDF* (Aizawa 2003). The idea is quite simple: weight the word count by the inverse of its frequency over different documents. A word that appears on every document will have a small weight, as a rare word

will have a larger weight. It can be understood in two parts: the common locally part (word count in a single document, or a single incident description), or the *TF* part, and the global rarity (the rarity of the word usage over all documents), or the *IDF* part. For a document d, the TF term of a word t is its count $f_{t,d}$, as shown in equation 1:

$$\text{TF}(t, d) = f_{t,d} \tag{1}$$

For the same word t, the IDF term over all N documents is:

$$\text{IDF}(t) = \log\left(\frac{N}{1 + n_t}\right) \tag{2}$$

where n_t is the number of documents the word t appears in. Then, the TF_IDF is calculated as:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \cdot \text{IDF}(t) = f_{t,d} \cdot \log\left(\frac{N}{1 + n_t}\right) \tag{3}$$

MODELING

After applying the TF-IDF in the data, we end up with a table containing weights of each word for each one of the entries (injury reports). In the total, after some filtering, we have 267 words (that are our features). However, the target (injury classification), contains 168 classes, which is very high, and correctly classifying them will be challenging. There are some of the injuries that happen just once in the whole data set, while *fractures* and *amputations* (the most common injury types), happen more than ten thousand times. This imbalance in the data is too uneven, and we decided to remove the less frequent injuries. Our criterion was to remove classes with low counts, but still keep 95% of the data. This happened when we chose classes with frequencies equal or larger than 100, ending up with a total of 32 classes (Table 2).

Table 2: All injuries classes used for modeling and their frequency.

Injury	Abbreviation	Frequency
Fractures	Frct	12547
Amputations	Ampt	11134
Soreness, pain, hurt-nonspecified injury	Sphi	3420
Cuts, lacerations	Ctsl	2068
Traumatic injuries and disorders, unspecified	Tiadu	1249
Intracranial injuries, unspecified	Iniu	897
Crushing injuries	Crsi	847
Heat (thermal) burns, unspecified	Htbu	764
Internal injuries to organs and blood vessels of the trunk	Iitoabvott	684
Puncture wounds, except gunshot wounds	Pwegw	631
Fractures and other injuries, n.e.c.	Faoin	456
Effects of heat and light, unspecified	Eohalu	367

Table 2: All injuries classes used for modeling and their frequency. *(continued)*

Injury	Abbreviation	Frequency
Electrical burns, unspecified	Elbu	365
Bruises, contusions	Brsc	350
Effects of heat and light, n.e.c.	Eohaln	344
Second degree heat (thermal) burns	Sdhtb	323
Concussions	Cncs	308
Electrocutions, electric shocks	Eles	298
Avulsions, enucleations	Avle	256
Chemical burns and corrosions, unspecified	Cbacu	248
Fractures and dislocations	Frad	247
Poisoning, toxic, noxious, or allergenic effect, unspecified	Ptnoaeu	230
Third or fourth degree heat (thermal) burns	Tofdhtb	223
Heat exhaustion, prostration	Htep	178
Cerebral and other intracranial hemorrhages	Caoih	160
Dislocation of joints	Dsoj	148
Other respiratory system symptoms-toxic, noxious, or allergenic effect	Orssnoae	148
Major tears to muscles, tendons, ligaments	Mttmtl	131
Fractures (except skull fractures) and concussions	Fesfac	114
Hernias due to traumatic incidents	Hdtti	113
Multiple effects of heat and light	Meohal	109
Gunshot wounds	Gnsw	107

Predictions using the TF-IDF weights

The idea is to use the TF-IDF weights to predict the injury classes of Table 1. But our goal is to focus on the predictions in the Oil & Gas industry, by separating it as the *testing set*, and use all the other industries to *train* a classification model. The training and test sets contain the remaining 32 classes of the original data.

Our first step to find a good model is to define a baseline. We want to make sure that the models tested are better than random guesses, so we trained a dummy classifier model, which is a pseudo-random guess model. The "pseudo" derives from the classifier having its randomness weighted by the classes{'} imbalanced distribution (as shown in Table 2). Figure 10 shows the confusion (a crossplot with normalized counts of each observation real classification versus its prediction) for the dummy classifier. A perfect model would generate a confusion matrix with diagonal of 1 and the other elements 0. The dummy classifier shows no pattern in the confusion matrix, and its accuracy is 19.4%. However, any model, in an imbalanced data set, will have higher accuracy if it predicts the most frequent classes more often. A better metric in this case is the *balanced accuracy*, where the final value is the sum of each class accuracy multiplied by a weight inversely proportional to its frequency. And the dummy classifier model balanced accuracy is only 3.4%.

Now that we know the "worst" a model can do, let's start to train classification models. As the data has imbalanced classes, we will rather train models that behave better in these situations. The first model we trained was an ensemble classification method called extremely randomized trees (Geurts, Ernst, and Wehenkel 2006), which is a step further of *random forests* (Breiman 2001) relative to the randomness. Random forests divide the train set into several subgroups and train a different decision tree for each one. In the end

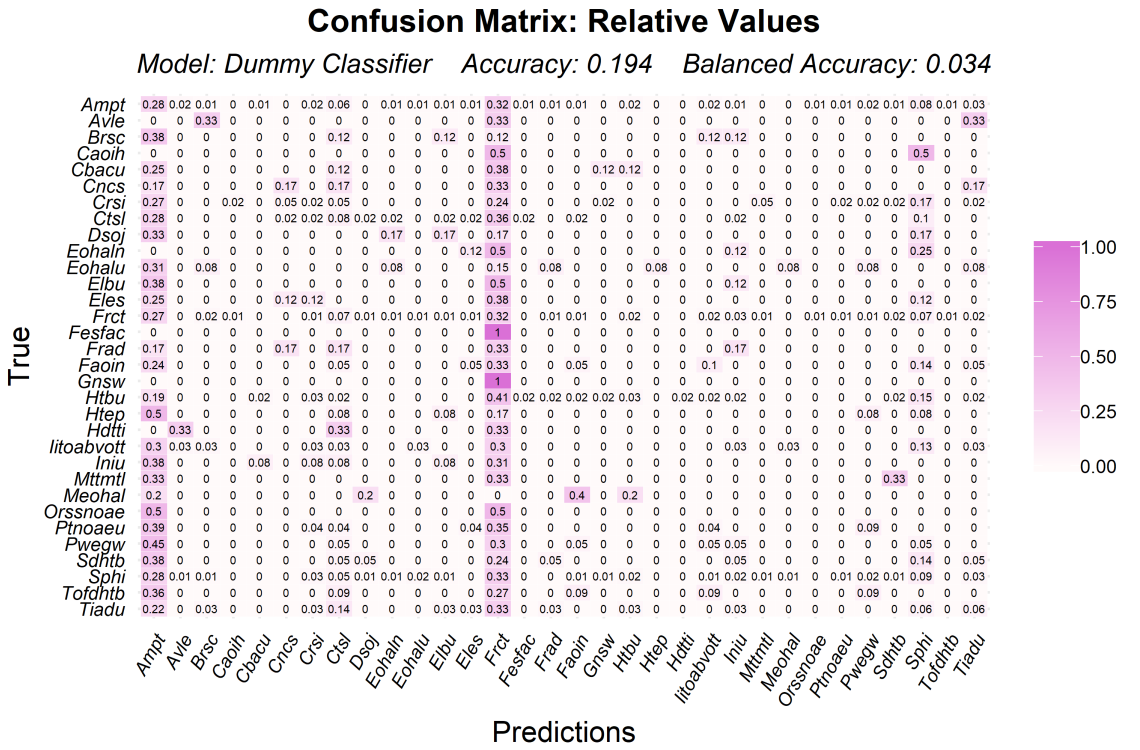

Figure 10: Confusion matrix for the dummy classifier model.

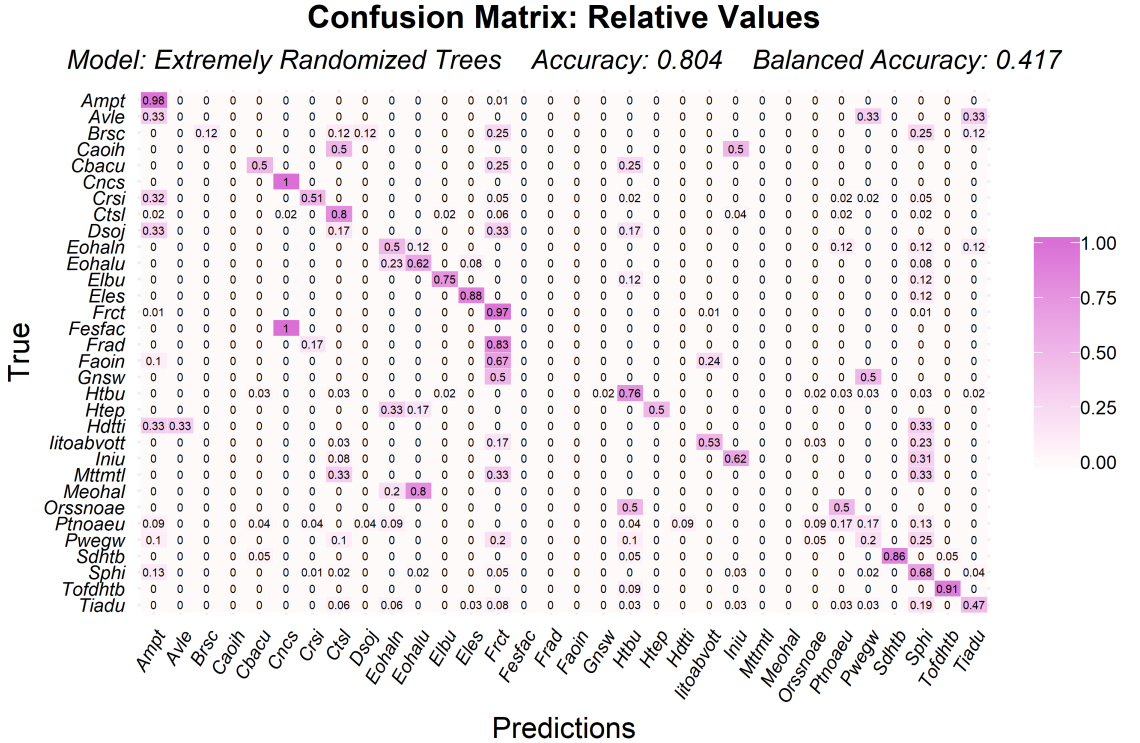

Figure 11: Confusion matrix for the random forest model.

they apply a vote system over all the trees to compute the prediction. The extremely randomized trees also has random splits inside each decision tree (the criterion to split a leaf into other leaves), which provides a more robust model, with lower variance (how much a

model change when trained on different data sets), but that can slightly increase the bias (the ability of the model to match training set. The higher the bias is, lower is the match). Figure 11 shows the confusion matrix for the extremely randomize trees model. It is apparently accurate, with accuracy of 80.4%. However, for the oil & gas industry, amputations and fractures (only) are the most frequent classes, representing 59% of the total of the testing set. So, accurately classifying these two classes will push the accuracy up (look at the balanced accuracy at 41.7%). But the model also worked well for other classes as well, or the accuracy wouldn't reach 80%, like *concussions* (CNCS), for example, whose classification was 100% accurate. However, the model failed when classes were similar, as the in the case of *fractures* family, which tended to be classified simply as *fractures* (FRCT), with the exception of *fractures (except skull fractures) and concussions* (FESFAC), that were sent 100% to concussions. A curious case are the *effects of heat and light, unspecified* (EOHALU) and *multiple effects of heat and light* (MEOHAL). Those look to be **very** similar classes (if not the same), and they are easily misjudged by the model. Maybe another pre-processing of the data should be applied to unify classes which are too similar.

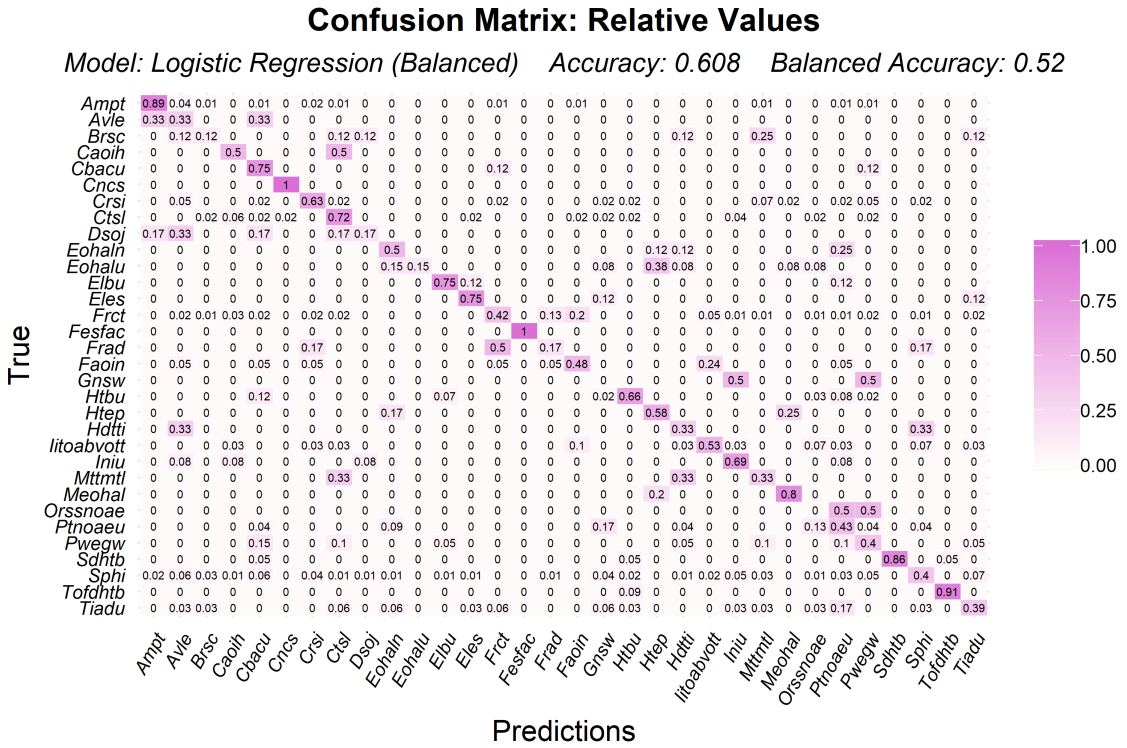

Figure 12: Confusion matrix for the logistic regression model with balanced class weights.

Another model tested was the logistic regression (Böhning 1992), which gave us the ability to balance the classes during modeling by applying weights that are proportional to the inverse of their frequency. Figure 12 shows its confusion matrix and it appears to have a better defined diagonal. The accuracy decreased to 60.8% when compared to the extremely randomized trees model, but the balanced accuracy increased to 52%, meaning the logistic regression model is working better for less frequent classes, but is failing to predict properly the most common classes. It is clear when we look at *fractures* (FRCT), where its predictions were separated into several different classes (but around 65% are still inside the *fractures family*). However, cases like *concussions* (CNCS) and *fractures*

(except skull fractures) and concussions (FESFAC) are not being confused by the model. Another curiosity are that *gunshot wounds* (GNSW) are being totally missclassified: half went to *intracranial injuries, unspecified* (INIU) and half to *puncture wounds, except gunshot wounds* (PWEGW), exactly to what it should not be classified as. Probably the word *wound* played an important role for this misclassification. Also *other respiratory system symptoms-toxic, noxious, or allergenic effect* (ORSSNOAE) is easily confused with *poisoning, toxic, noxious, or allergenic effect, unspecified* (PTNOAEU), as both look to be the same class in a level.

Predictions using binary input features

For a new approach, instead of using the TF-IDF weights for each word as the numerical representation of them, we will use a binary representation. This means that it will not matter how many times a word appears in a description, it will filled as 1 if present, and 0 otherwise. We want to see if the binary approach balances the features importance and avoid some of the misclassifications.

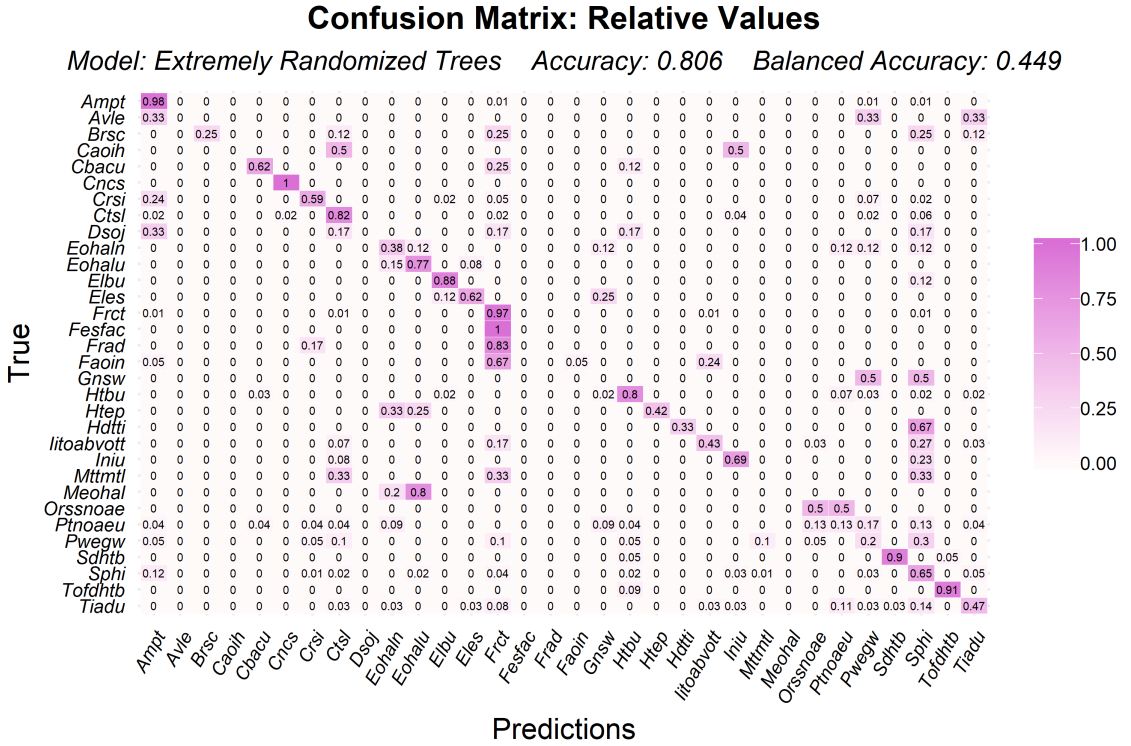

Figure 13: Confusion matrix for the random forest model for binary features.

Figure 13 is the confusion matrix generated using the extremely randomized trees model with the new binary strategy. The accuracy remained almost the same, but the balanced accuracy increased to 44.9%, as the model started to differentiate the least frequent classes better. But the same analysis for the previous approximation is still valid here in an overall scenario. We still need to improve the classification for this model for similar classes.

Going back to the balanced logistic regression, its confusion matrix (Figure 14) shows

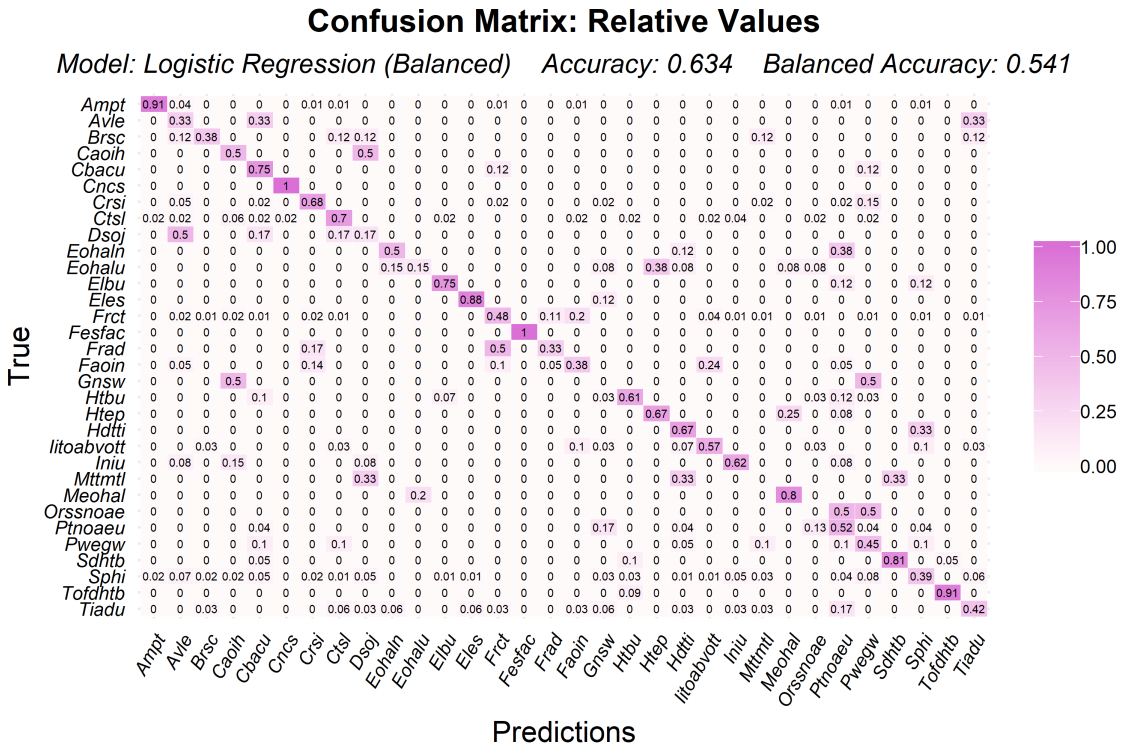

Figure 14: Confusion matrix for the logistic regression model with balanced class weights and binary features.

significant improvements for both the accuracy (63.4%) and balanced accuracy (54.1%). *Amputations* (AMPT) and *fractures* (FRCT) have improved accuracy. Apparently, the logistic regression is a more robust model in an overall scenario.

Changing to the binary input features approximation improved the performance of both models. However, the *extremely randomized trees* model works better for the most frequent classes (more specifically *fractures*) while the *logistic regression* works much better for the least frequent classes (actually, for the most frequent classes, only *fractures* showed to be challenging for this model). We actually believe that the logistic regression can be the most accurate model if we work better with the features, instead of using only single words as features, we could implement the pairs of sequential words (bi-gran) from descriptions.

Another observation to keep in mind is that that we removed the reports from the oil & gas industry to train the models, then tested their performance on the oil & gas industry data. It would be interesting in a continuing project to do the analysis of the type of language used in different industries. Are the descriptions in the oil & gas industry data the same as for others industries? Or does each industry have its own terminology, and models are failing to recognize this? We believe these are questions that need to be answered in future projects.

CONCLUSIONS

In this work, we studied severe injury reports from the *Occupational Safety and Health Administration of the United States Department of Labor*, where we analyzed the incidents

locations, and most frequent injuries by location and type, and worked on a project to predict the injury classification from its description, with the goal to create a model that can standardize the classification for the government department.

One of the main steps of the project was the "conversion" of the descriptions into numerical variables by using the count of each word in the descriptions as initial features, and then removing stop words (common words) using the TF-IDF algorithm. It creates a matrix of weights where the columns are the words and the rows are the observations (injury reports). The data are then separated into testing (all the oil & gas industry reports) and training (the remaining industries) sets.

The trained models presented different performances for the imbalanced dataset. The *extremely randomized trees* have results of higher accuracy over the most frequent classes, more specifically for *fractures*, while the *logistic regression* showed a better overall performance, with higher accuracy for the least frequent classes, but failing to properly classify fractures.

By changing the input features from the TF-IDF weights to only binary values, both of the models showed improved performance. But we came to the conclusion that we need to improve more the feature engineering before modeling, and also understand better the language used by each industry and, whether it is an important characteristic of the data.

ACKNOWLEDGMENTS

The authors thank the sponsors of CREWES for continued support. This work was funded by CREWES industrial sponsors and NSERC (Natural Science and Engineering Research Council of Canada) through the grant CRDPJ 461179-13. We also thank Soane Mota dos Santos for all the knowledge share during very useful conversations.

REFERENCES

Aizawa, Akiko. 2003. "An Information-Theoretic Perspective of Tf-Idf Measures." *Information Processing & Management* 39 (1): 45–65.

Böhning, Dankmar. 1992. "Multinomial Logistic Regression Algorithm." *Annals of the Institute of Statistical Mathematics* 44 (1): 197–200. https://doi.org/10.1007/BF00048682.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Chokor, Abbas, Hariharan Naganathan, Wai K. Chong, and Mounir El Asmar. 2016. "Analyzing Arizona Osha Injury Reports Using Unsupervised Machine Learning." *Procedia Engineering* 145: 1588–93. https://doi.org/10.1016/j.proeng.2016.04.200.

Dublin, Sascha, Eric Baldwin, Rod L. Walker, Lee M. Christensen, Peter J. Haug, Michael L. Jackson, Jennifer C. Nelson, Jeffrey Ferraro, David Carrell, and Wendy W. Chapman. 2013. "Natural Language Processing to Identify Pneumonia from Radiology Reports." *Pharmacoepidemiology and Drug Safety* 22 (8): 834–41. https://doi.org/10.1002/pds.3418.

Geurts, Pierre, Damien Ernst, and Louis Wehenkel. 2006. "Extremely Randomized Trees." *Machine Learning* 63 (1): 3–42. https://doi.org/10.1007/s10994-006-6226-1.

Shahana, P.H., and Bini Omman. 2015. "Evaluation of Features on Sentimental Analysis." *Procedia Computer Science* 46: 1585–92.

Tixier, Antoine J.-P., Matthew R. Hallowell, Balaji Rajagopalan, and Dean Bowman. 2016. "Automated Content Analysis for Construction Safety: A Natural Language Processing System to Extract Precursors and

Outcomes from Unstructured Injury Reports." *Automation in Construction* 62: 45–56. https://doi.org/10.1016/j.autcon.2015.11.001.

Weber, Dean. 2002. Interactive user interface using speech recognition and natural language processing. Patent Number: US6499013B1, issued 2002.

Yetisgen-Yildiz, Meliha, Cosmin Bejan, and Mark Wurfel. 2013. "Identification of Patients with Acute Lung Injury from Free-Text Chest X-Ray Reports," August, 10–17.