Disaster Tweets Classification

Alina Skrylnik
alina.skrylnik@studenti.unipd.it

Brenda Eloísa Téllez Juárez
bredaeloisa.tellezjuarez@studenti.unipd.it

1. Introduction

Machine learning allows us to recognize and identify behavior patterns in the data, make predictions and thus be able to have solutions for different problems. In this project, we use these algorithms to detect and classify the disaster messages that appear on Twitter. For this purpose, we use the following models: KNN, logistic regression, Random Forest Classifier and MLP Classifier.

In the end, the model that has a better fit for the objective of this project is the Logistic Regression with an accuracy of 0.79.

2. Dataset

For the development, the Natural Language Processing with Disaster Tweets dataset was used, which is made up of 7613 tweets that must be classified as disaster or non-disaster. It consists of training set, which is split in train (6090 tweets) and validation sets (1523 tweets), and test set (3263 tweets). To process the data, we perform data cleaning for which we reduce the text, eliminate square brackets, hyperlinks, punctuation, and words that contain numbers.

In the dataset, we can see the most common words in the disaster tweets that are shown in figure 1.

Fig. 1. Keywords in 'disaster' tweets

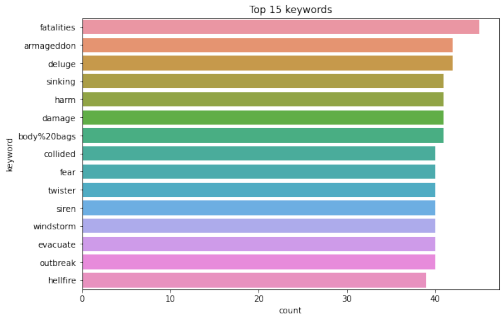

Figure 1. Source: Own elaboration with data from https://www.kaggle.com/c/nlp-getting-started

The places where the disaster is noticed most of all are: the USA, London, Canada, Nigeria, UK, India, Mumbai, Kenya, and the rest of the world.

Fig. 2. The most frequent places denoted in the disastrous tweets

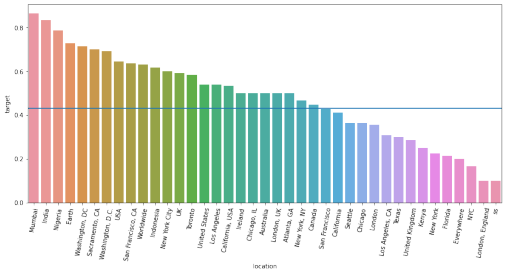

Figure 2. Source: Own elaboration with data from https://www.kaggle.com/c/nlp-getting-started

A Word Cloud was made to highlight the most important words:

Fig. 3. Clouds of words of disaster and non-disaster tweets

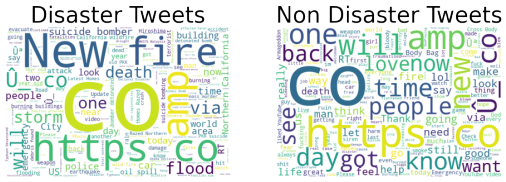

Figure 3. Source: Own elaboration with data from https://www.kaggle.com/c/nlp-getting-started

3. Models

To achieve the classification of the text we decided to perform these algorithms: K-Nearest Neighbor, Logistic regression, Random Forest Classifier and Neural Networks.

3.1. K-Nearest Neighbor

K-Nearest Neighbor is not a very complex algorithm that helps to realize and solve classification and regression problems, which search the most similar characteristics to implement the grouping. In our model the obtained accuracy is 0.72, we performed the GridSearchCV to obtain the best score which was K=3.

Fig. 4. K selection

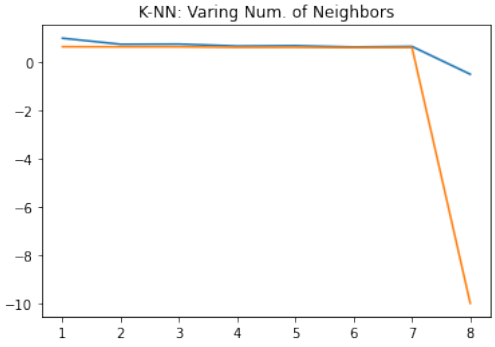

Figure 4. Source: Own elaboration with data from https://www.math.unipd.it/ dasan/disaster/

3.2. Logistic regression

Logistic regression uses discriminant analysis to perform the classification for categorical and predicted values. It is based on multivariate statistics. In the implementation of our model, the accuracy was 78%, using the best parameters with GridSearchCV and we calculated the Roc Curve and Roc Auc Curve to determine the performance of our classifier, so we can see that it has an excellent performance, since the Auc is 0.86.

Fig. 5. ROC AUC Curve

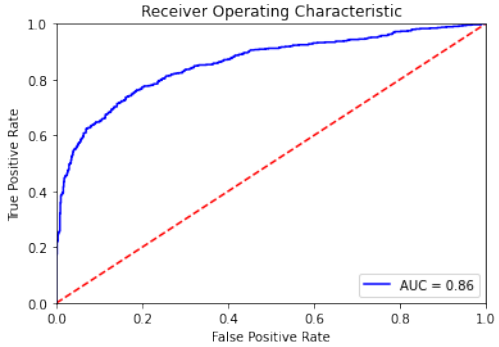

Figure 5. Source: Own elaboration with data from https://www.math.unipd.it/ dasan/disaster/

3.3. Random Forest Classifier

This model creates other random trees which take the most voted as the better tree classifier. In the case of our project, we implemented GridSearchCV to obtain the best parameters, which was 'gini', 'max depth=8', 'auto', and '200 estimators' with and accuracy of 0.62 .

3.4. Neural Networks

Finally we performed the Multi-Layer Perceptron which is composed of multiple perceptors in three layers of nodes. In our project the accuracy is 0.51, being the lowest.

4. Conclusion

- k-NN. Best parameters are: k=3. Class 0: 0.79, Class 1:0.55, accuracy: 0.72.
- Logistic Regression. Best parameters are: C = 0.44, penalty = 'l2'. Class 0: 0.84, Class 1: 0.72, accuracy: 0.78.
- Random Forest. Best parameters are: criterion = 'entropy', max_depth = 10, min_samples_split = 10. Class 0: 0.78, Class 1:0.28, accuracy: 0.67.
- Neural Network. Class 0: 0.74, Class 1:0 accuracy: 0.59.

In sum, the best model for classifying the disaster and no disaster tweets is the logistic regression which also presents the best accuracy over the rest of the models.

5. References

- Zhang, C., Pan, X., Huapeng, L., Gardiner, A., Sargent, I. (n.d.). A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification.
- Baoxun, X., Xiufeng, G., Yunming, Y., Jiefeng, C. (2012). An Improved Random Forest Classifier for Text Categorization. JOURNAL OF COMPUTERS.
- Breiman, L. (2004, 09 09). CONSISTENCY FOR A SIMPLE MODEL OF RANDOM FORESTS. BERKELEY.
- Kaggle. (n.d.). Retrieved from https://www.kaggle.com/c/nlp-getting-started scikit-learn. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html