Which techniques you have used while cleaning the data if you have cleaned it?

• Removed Stop Words

• Removed Unwanted Symbols

• Split by Whitespace and Removed Punctuation

• Removed Salaries "digits"

• Converted All Text To Onecase (Lowercase)

Why have you chosen this classifier?

I used Linear Support Vector Machine:

- Because it is widely regarded as one of the best text classification algorithms,
- I achieved a higher score over Naive Bayes with about 5% and over RandomForestClassifier with about 2%.
- SVMs is an algorithm that determines the best decision boundary between vectors that belong to a given group (or category) and vectors that do not belong to it,
- This means that in order to leverage the power of svm text classification, texts have to be transformed into vectors and I did that.

How do you deal with (Imbalance learning)?

I used the right evaluation metrics:

- Precision/Specificity: how many selected instances are relevant.
- Recall/Sensitivity: how many relevant instances are selected.
- F1 score: harmonic mean of precision and recall.

How can you extend the model to have better performance?

• Add more data for balancing classes.

• Trying to clean the data more and get out the useless and non meaningful words.

• Try Ensembles(combiine weak model to get better results)

How do you evaluate your model?

I evaluated using:

- Precision
- Recall
- F1 score

due to the imbalancement of the data classes, so if accuracy is used to measure the goodness of a model, a model which classifies all testing samples into “0” will have an excellent accuracy (99.8%), but obviously, this model won’t provide any valuable information for us.

What are the limitations of your methodology or Where does your approach fail?

• Takes long training time

• so it will be bad on large datasets

Refrences:

1- https://towardsdatascience.com/multi-class-text-classification-model-comparison-and-selection-5eb066197568

2- https://monkeylearn.com/text-classification-support-vector-machines-svm/

3- https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html

4- https://www.researchgate.net/publication/2522390_Improving_Multiclass_Text_Classification_with_the_Support_Vector_Machine

5- https://towardsdatascience.com/machine-learning-multiclass-classification-with-imbalanced-data-set-29f6a177c1a

6- https://towardsdatascience.com/simplify-your-dataset-cleaning-with-pandas-75951b23568e

7- https://www.kaggle.com/pamin2222/tf-idf-svm-exploration