Multiclass Text Classification

Data Scientist Application Exercise

Andrés Troiano

July 6, 2021

1 Business Understanding

Konfío is a Mexican fintech company that offers credit for small and medium-sized businesses. During the evaluation process for potential clients, applicants are asked to provide a short written use of proceeds. That is, a description of how they are planning to spend the capital. Understanding these intentions is very important for Konfío, and the company would like to be able to classify these descriptions into different categories, e.g. to pay debt, to buy new equipment, or for payroll.

The company has tested two approaches to solve this problem: Machine Learning (SVC, Logistic Regression, Random Forest, Multinomial Naive Bayes) and Deep Learning (CNN). The best performing algorithm was Logistic Regression, with a loss score of 0.068 and an average precision of 0.67. The goal is to provide an alternative solution that hopefully performs better, while cycling through the Cross Industry Standard Process for Data Mining (CRISP-DM).

This work will also cover two approaches: Machine Learning and Deep Learning, although in this case a recurrent neural network will be used instead of a convolutional one.

2 Data Understanding

The dataset provided has 11 columns. The first one contains the uses of proceeds. That is, each row has the description of the intentions of one applicant. The remaining columns contain the provided labels for the texts. A description of each category can be found on section 7:

The dataset has one column for each of these categories, with a value of 1 every time the description falls into said category, and 0 every other time. More than one labels are allowed. The dataset totals 6679 uses of proceeds. The first five rows are shown in figure

1:

	motivos	crec	cred	equ	inic	inv	mkt	no	renta	sueldo	temp
0	Crear un departamento de ventas e inversión a ...	0	0	0	0	0	1	0	0	0	0.0
1	establecerme en un local y agregar materia pri...	0	0	0	0	1	0	0	1	0	0.0
2	Compra de equipo e incrementar inventario	0	0	1	0	1	0	0	0	0	0.0
3	Invertir en crecimiento de flotilla de unidade...	0	0	1	0	0	0	0	0	0	0.0
4	Para comprar mercancía y comprar lonas nuevas	0	0	0	0	1	0	0	0	0	0.0

Figure 1: *First 5 rows of the dataset.*

Representation of each class: Figure 2 shows the class distribution, which as we can see is imbalanced. This is often the case in real business problems, and means that accuracy is a poor choice as a performance metric. The reason is that accuracy gives high scores to models which just predict the most frequent class. Precision, recall, and F1 score might be a better choice in this case. Certain applications such as fraud detection require that the dataset be artificially balanced. However in this case it is possible that the majority classes are of the greatest interest, so as a first order approximation the class distribution was kept as is.

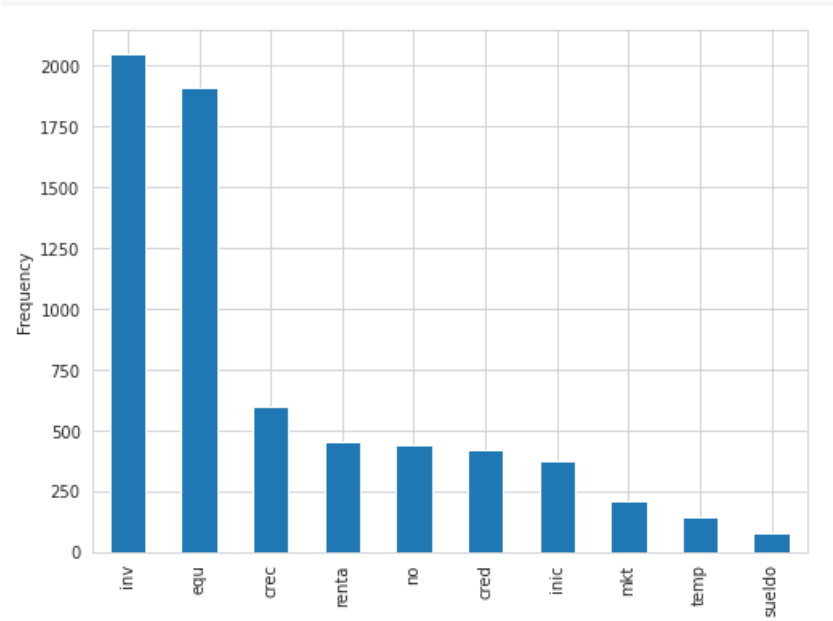

Figure 2: *Class distribution.*

Potential problems:

- Spelling mistakes (e.g. Ampliaciin, mas.habitaciones)
- Inadequate labels (e.g. "Compra de inmobiliario de oficina" labeled as "renta")
- Some proceeds dealt on more than one category but had only one label.

These mistakes could harm the ability to make predictions, because the models would learn the wrong labels. Deep Learning algorithms in particular are very robust to random errors in the training set, provided the number of errors is small. However, if these errors are systematic, the model could learn to misclassify. However, before dedicating time and effort to correct these mistakes, it is important to estimate whether this would make a significant improvement in performance. This requires a thorough error analysis that exceeds the deadline for this exercise.

3 Data Preparation

Data cleaning: The variable "temp" had 6 missing values, which were imputed using the median. Another option could have been to drop those rows, but since most of them did have other labels, it was considered best to keep them. Also this variable was of type *float* when it should be *int*, so it was converted accordingly.

3.1 Text pre-processing

Text pre-processing was different for the Machine Learning and Deep Learning approaches.

Machine Learning: At the first iteration through the CRISP-DM cycle, the priority was to build a model that works, even if it didn't perform too well. At this stage it was considered of paramount importance to have something to work on and build upon, that can later be improved.

The selected Machine Learning classifiers work under the assumption that each use of proceeds is assigned to one and only one category. This is a significant simplification of the problem worth revisiting in the future, because in general the proceeds had more than one label. The criteria for label selection in those cases was to keep the first one from left to right.

Spanish stop words were removed. In order to extract features from the text, the Bag of Words model was used, which considers the presence and frequency of each word, but ignores the order in which they appeared. This seems appropriate because the texts were short. Text was vectorized using *tf-idf*.

Deep Learning: For the RNN, all text was converted to lower case. The following symbols were replaced by space:

```
/(){}[]|@,;
```

All symbols that are not digits, letters, space, or the following, were removed: #+-

Spanish stop words were removed, digits in text were removed (e.g. mercade5ria replaced by mercaderia). Text was tokenized using Keras's Tokenizer class.

4 Modeling

Machine Learning: The following classifiers were tested and compared: Multinomial Naive Bayes, Random Forest, Linear SVC, and Logistic Regression (using the multinomial approach, and cross-entropy as the loss function). Hyperparameter tuning was not performed due to time constraints.

Deep Learning: Recurrent Neural Networks (RNNs) are ideal for sequential data such as text. For this task the LSTM model was chosen (Long Short-Term Memory). The first layer is the embedding layer that uses 100 length vectors to represent each word. The second layer is SpatialDropout1D, which performs variational dropout. The third layer is the LSTM layer with 100 memory units. The output layer (dense) creates 10 output values, one for each class. Activation function is softmax for multi-class classification. Because this is a multi-class classification problem, categorical crossentropy was used as the loss function.

5 Evaluation

5.1 Machine Learning

K-fold cross validation was performed, using 5 stratified folds. Figures 3, 4, 5 and 6 show accuracy, precision, recall and F1 score respectively. Table 1 shows the average for every metric across the 5 folds.

Model	Accuracy	Precision	Recall	F1 score
Logistic Regression	0.69	0.65	0.48	0.52
Linear SVC	0.69	0.61	0.55	0.57
Multinomial NB	0.61	0.60	0.34	0.37
Random Forest	0.46	0.12	0.15	0.12

Table 1: *Average metrics across K-folds for each model.*

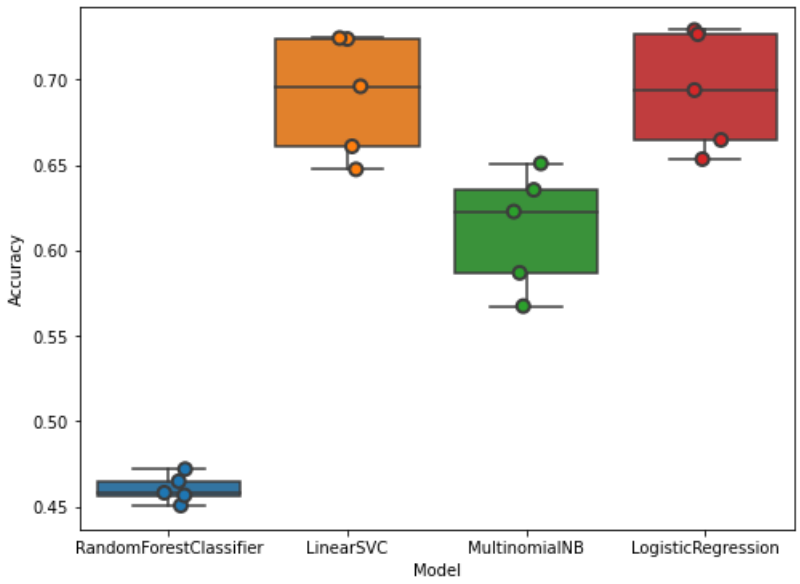

Figure 3: *Accuracy of the different ML classifiers during cross validation.*

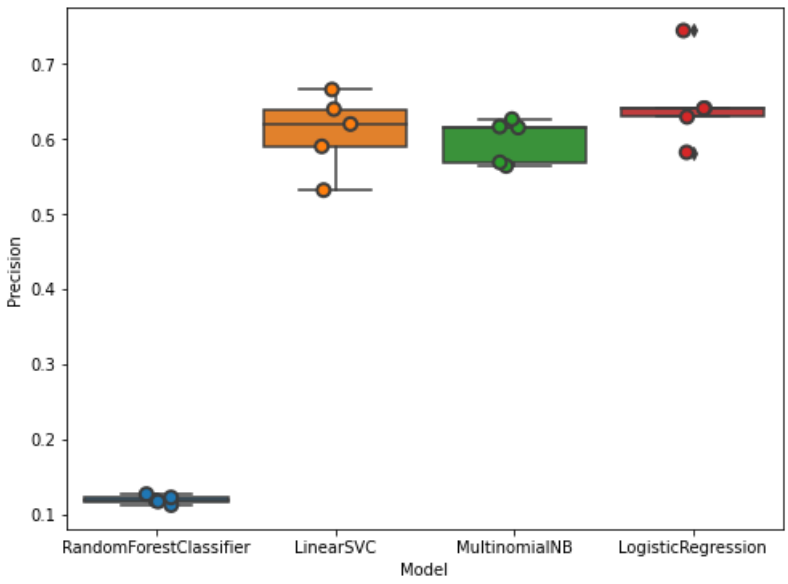

Figure 4: *Precision of the different ML classifiers during cross validation.*

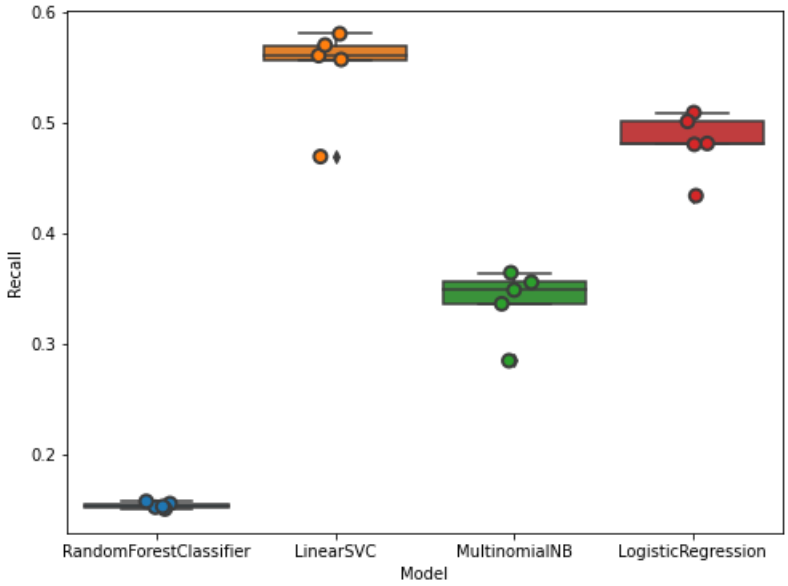

Figure 5: *Recall of the different ML classifiers during cross validation.*

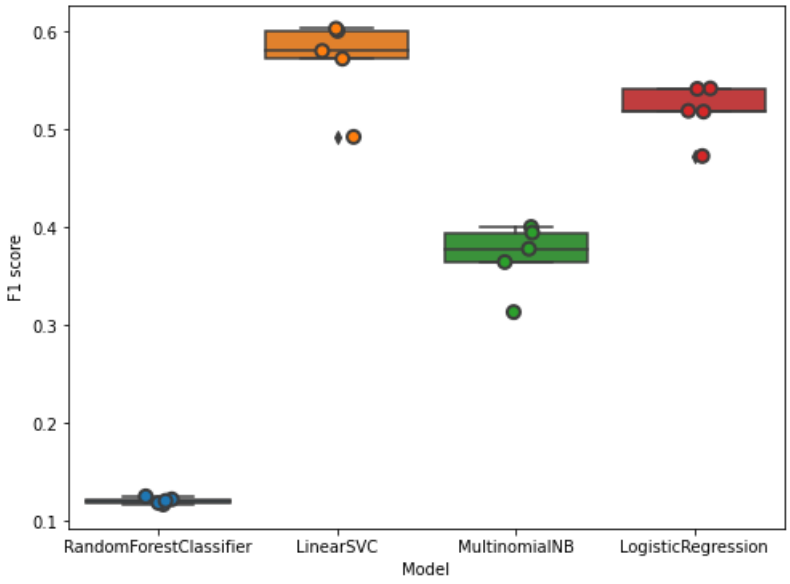

Figure 6: *F1 score of the different ML classifiers during cross validation.*

The best classifier was Logistic Regression with the multinomial approach, as measured by precision, achieving a score of 0.65. The *lbfgs* solver was initially chosen, but the model failed to converge, even though the results were stable between iterations. However, switching to *newton-cg* solver fixed the issue.

When training the Logistic model, warnings showed that not all labels were included in the predictions, i.e. there were some labels in the test set that the classifier never predicted. This happens when an imbalanced dataset is randomly split. However, stratified splitting did not fix the issue. This requires further investigation, but it would exceed the deadline.

Figure 7 shows the confusion matrix for the Logistic Regression model.

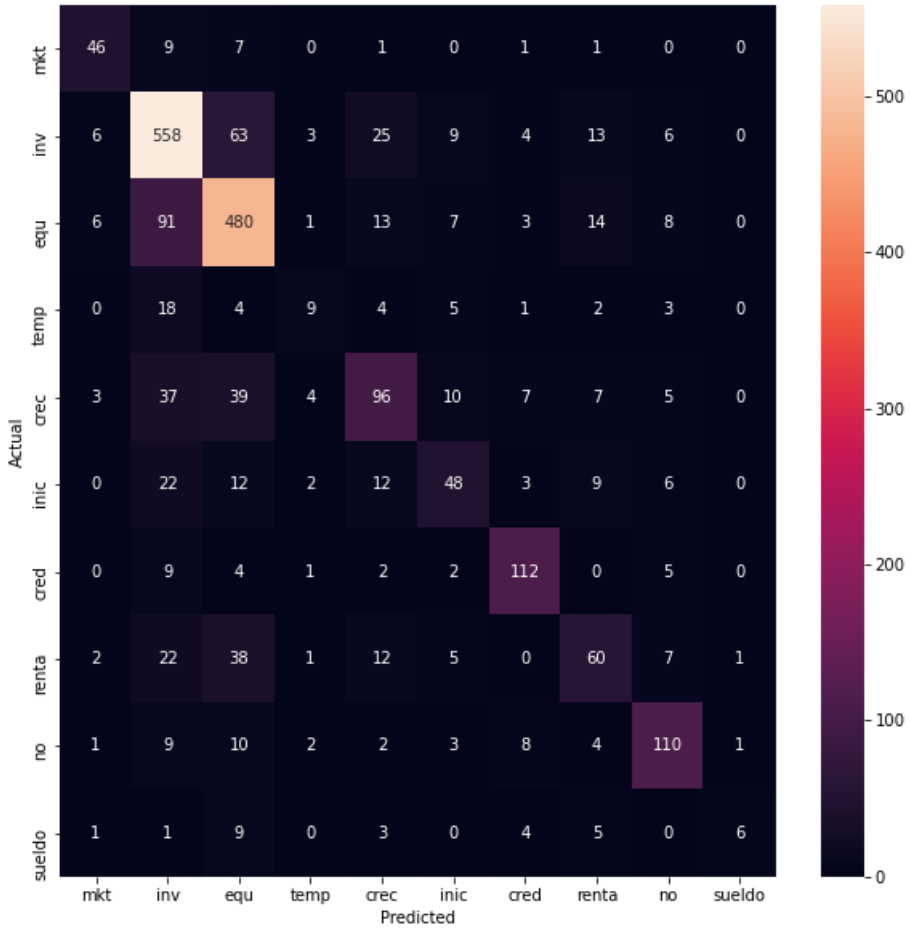

Figure 7: *Confusion matrix for Logistic Regression.*

Figure 7 shows that the classifier got the vast majority of predictions right (diagonal). However there are a number of misclassifications, and it is worth investigating. Figure 8 shows samples of text classified as *inv* when their actual label was *equ*.

For instance, rows 651 and 2643 were classified as *inv*, when their true label was *equ*.

'equ' predicted as 'inv' : 91 examples.

	category	motivos
444	equ	invirtiendo el dinero en la adquisición de equ...
651	equ	inversion en equipo y material para reestablec...
2643	equ	Compra de insumos para proyecto, materiales y ...
2858	equ	PARA COMPRAR MÁS MATERIA PRIMA Y PODER SURTIR ...
2753	equ	Para compra de material e infraestructura

Figure 8: *Misclassification samples.*

Upon inspection we find that in the original dataset they had both labels. This confirms that having kept only one label for each text hurts performance to some degree. In other cases predictions didn't match the original label, however the text did touch on both the real and the predicted topic. This is an example of mistakes caused by limitations in the original labeling.

5.2 Deep Learning

Table 2 shows loss, accuracy, precision and recall achieved by the LSTM model (F1 score is not available since training was performed in batches). As we can see precision is 0.73, an improvement from 0.67 achieved by the Logistic Model in previous works.

Loss	1.11
Accuracy	0.68
Precision	0.73
Recall	0.64

Table 2: *Performance metrics for the LSTM model.*

Figures 9, 10, 11 and 12 show the evolution of the different metrics over the epochs. These figures show strong overfitting, as evidenced by an increase in the loss, and a markedly inferior performance on the test set compared to the training set. One possible cause is that the training and test splits are not statistically equivalent. To address this issue, stratification was performed. An improvement was observed, but overfitting wasn't completely eliminated. Further recommendations are to add regularization, and to lower the capacity of the model to memorize the training data. The second approach will cause the model to focus on the relevant patterns in the training data, which results in better generalization. However these solutions exceed the deadline for this exercise.

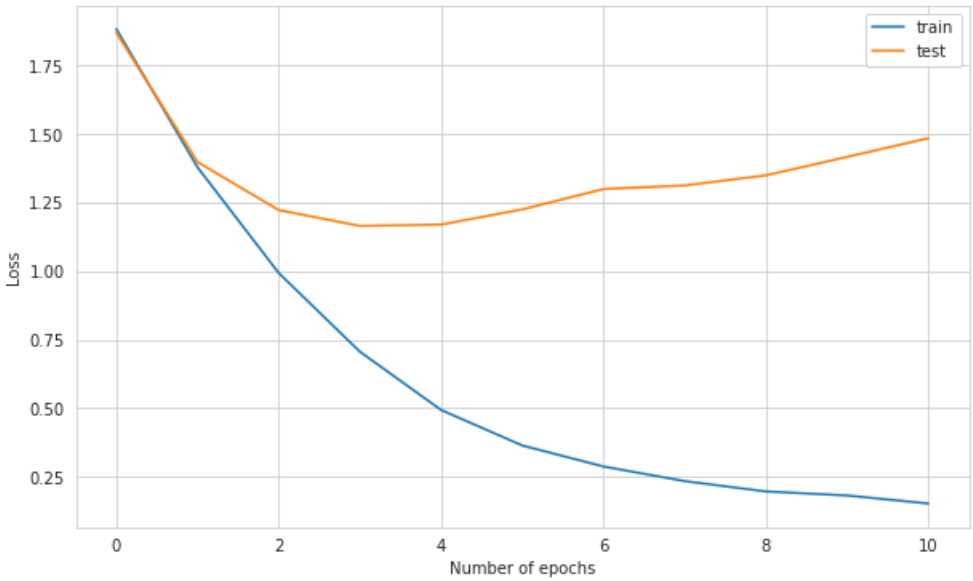

Figure 9: *Loss curve.*

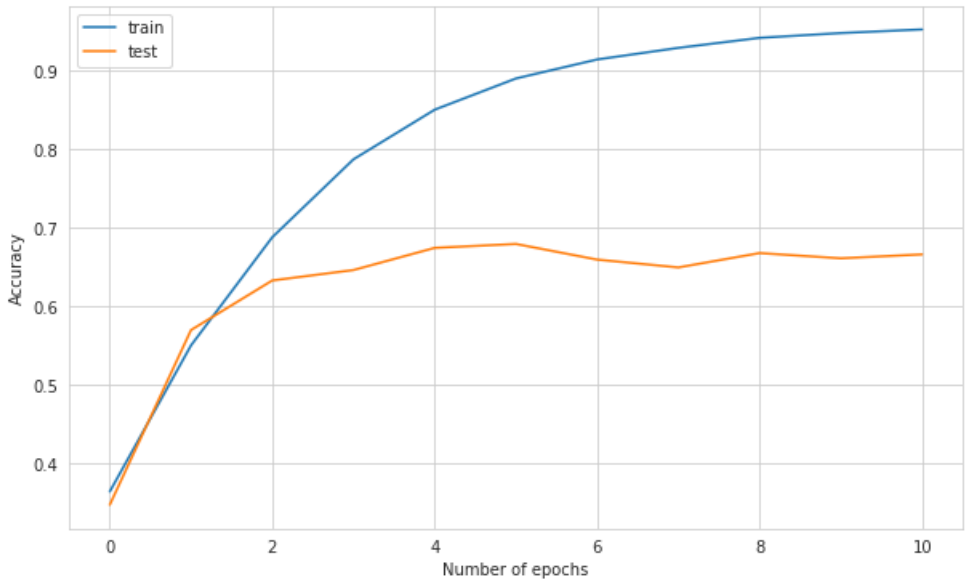

Figure 10: *Accuracy curve.*

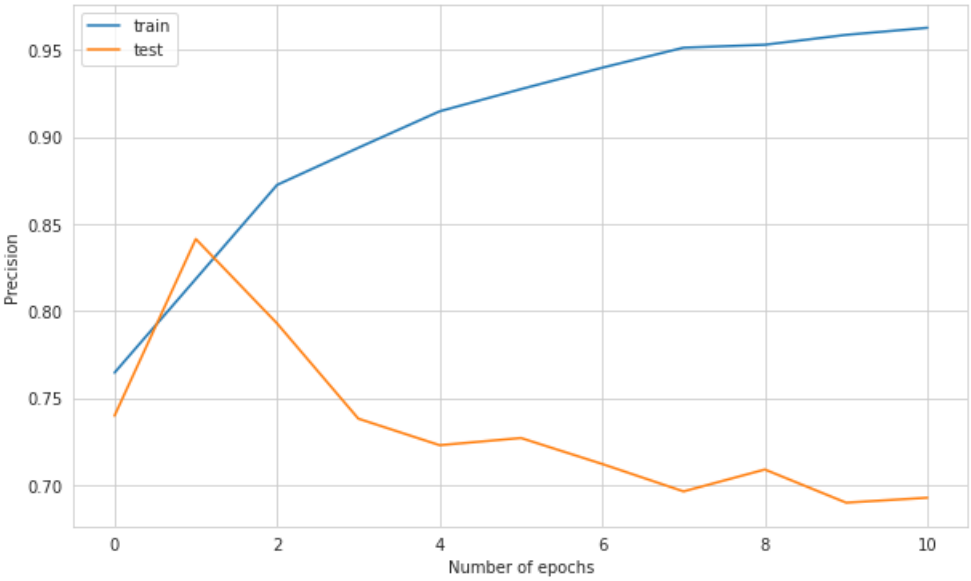

Figure 11: *Precision curve.*

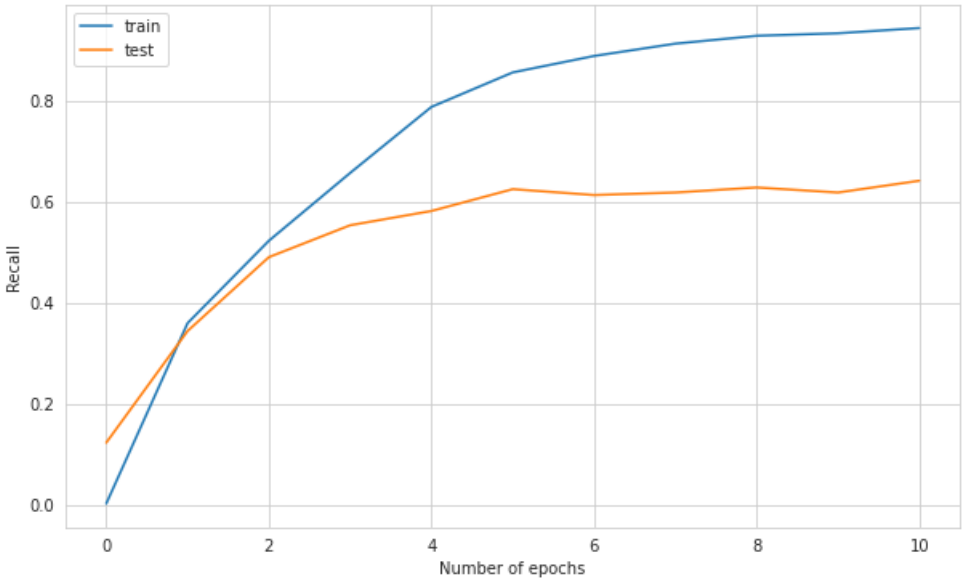

Figure 12: *Recall curve.*

6 Conclusion

Two main approaches were tested: traditional Machine Learning classifiers (Multinomial Naive Bayes, Random Forest, Linear SVC, and Logistic Regression), and Deep Learning (RNN using the LSTM model). The best performing model in terms of precision was the RNN, achieving a score of 0.73, thus improving the metric from previous works. However strong overfitting is observed. Recommendations to fix this issue are to balance the dataset, hyperparameter tuning, adding regularization, and trying different alternatives for text encoding. Time constraints prevented from further exploring these ideas in the present work.

7 Appendix

Description of the 10 categories:

- crec: insinuation of a strategy for growth, even if no specifics were provided.
- cred: debt payments.
- equ: purchasing equipment.
- inic: kickstarting the business.
- inv: purchasing inventory.
- mkt: marketing.
- no: not destined to working capital. In practice this often means personal or private expenses, like medical treatments or schooling.
- renta: pay the rent.
- sueldo: payroll.
- temp: seasonal expenses.