Multiclass Text Classification

Data Scientist Application Exercise

Andrés Troiano

July 6, 2021

1 Business Understanding

Konfío is a Mexican fintech company that provides funding 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, so the company would like to be able to classify these descriptions in different categories. Some examples are: to pay debt, to buy new equipment, or for payroll.

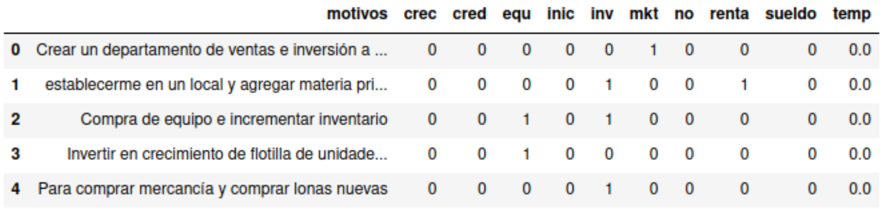
The company has tested two approaches to solve this problem: Machine Learning (SVC, Logistic Regression, Random Forest, Multinomial Naive Bayes) and Deep Learning (CNN). The best performing algorithm was Logistic Regression, with a loss score of 0.068 and an average precision of 0.67. The goal is to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

Similarly, this work will cover two approaches: machine learning algorithms, and deep learning, although in this case a recurrent neural network architecture will be used instead of a CNN.

2 Data Understanding

The dataset provided is made of 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 this 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. F1 score would be a better choice in this case. (También está la lectura de que las clases mayoritarias pueden ser las que más nos interesan.)

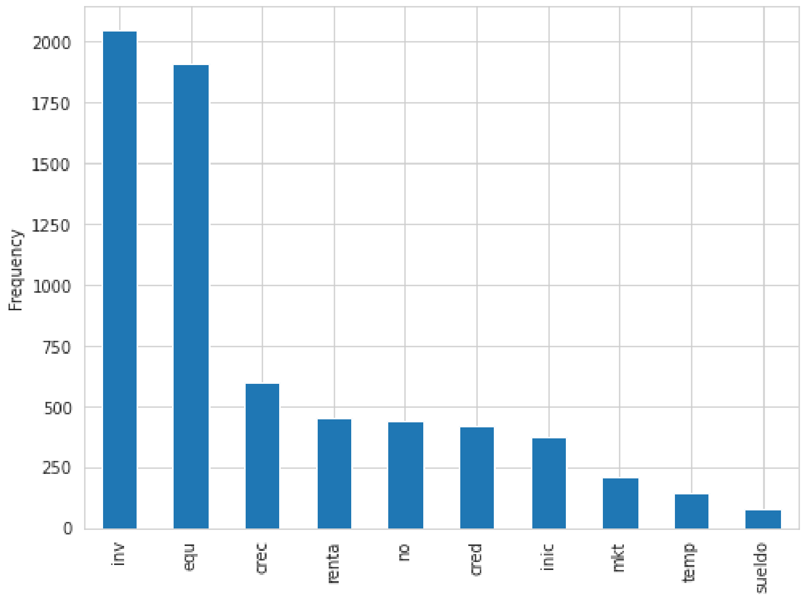


Figure 2: *Class distribution.*

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. 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 (Describir cómo se hace)

Deep Learning algorithms 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.

3 Data Preparation

Data cleaning: The variable "temp" had 6 missing values, which were imputed using the median (0). 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 vas of type float when it should be int, so it was converted accordingly.

3.1 Text pre-processing

Text pre-processing was slightly 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 doesn'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 machine learning classifiers used 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 where more than one was present, was to keep the first one from left to right.

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 (Poner la distribución de largos). 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: A comparison of different classifiers was carried out, comprising Multinomial Naive Bayes, Random Forest, Linear SVC and Logistic Regression.

Deep Learning: We used the LSTM model (Long Short-Term Memory), which is a Recurrent Neural Network architecture.

Alguna figura de Andrew? El dibujito del modelo?

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 it is a multi-class classification problem, categorical crossentropy is used as the loss function.

Por qué elegí estos modelos

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	0.694	0.647	0.481	0.518
Linear SVC	0.691	0.609	0.548	0.569
Multinomial NB	0.613	0.598	0.338	0.370
Random Forest	0.460	0.119	0.154	0.121

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

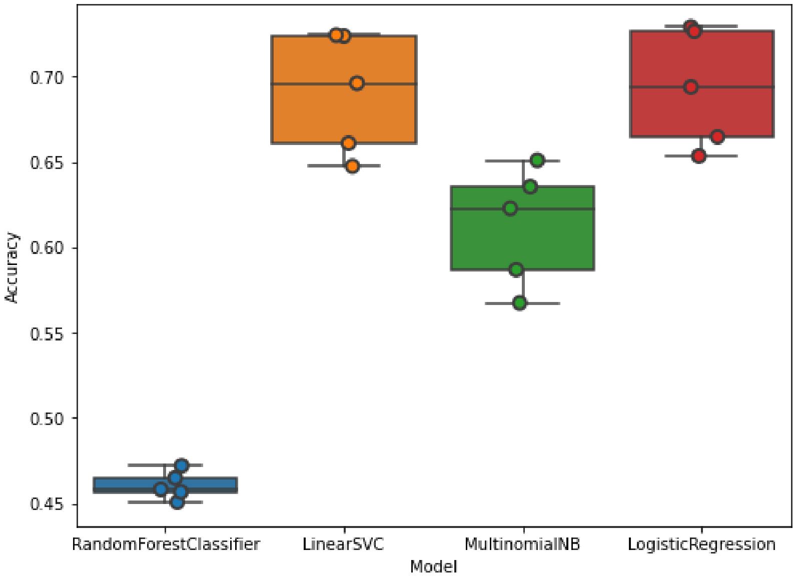


Figure 3: *Accuracy of the different ML classifiers during cross validation.*

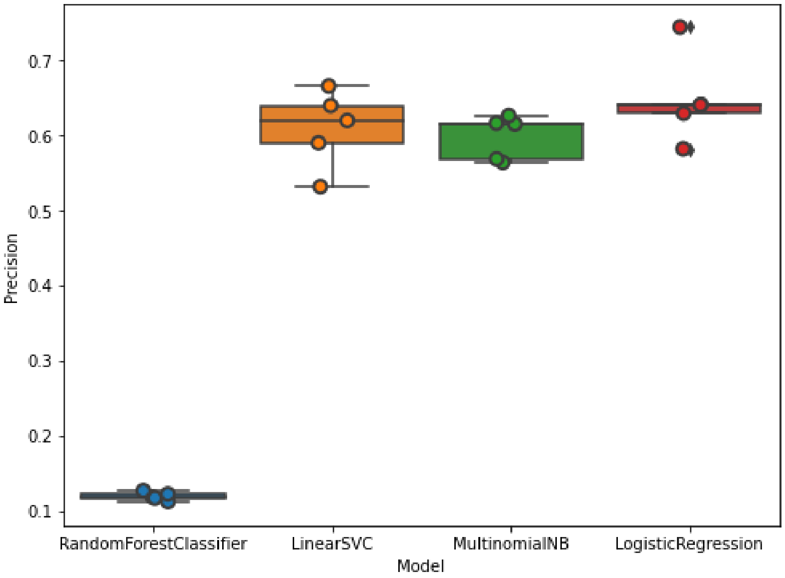


Figure 4: *Precision of the different ML classifiers during cross validation.*

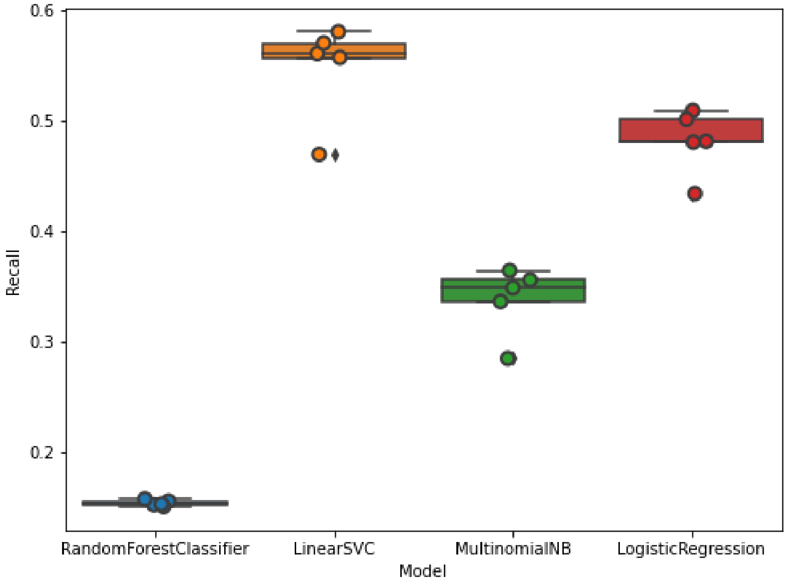


Figure 5: *Recall of the different ML classifiers during cross validation.*

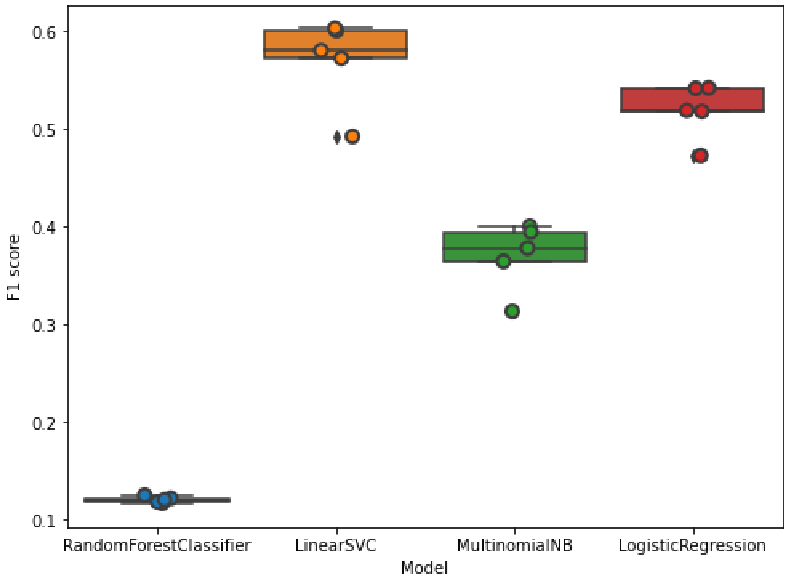
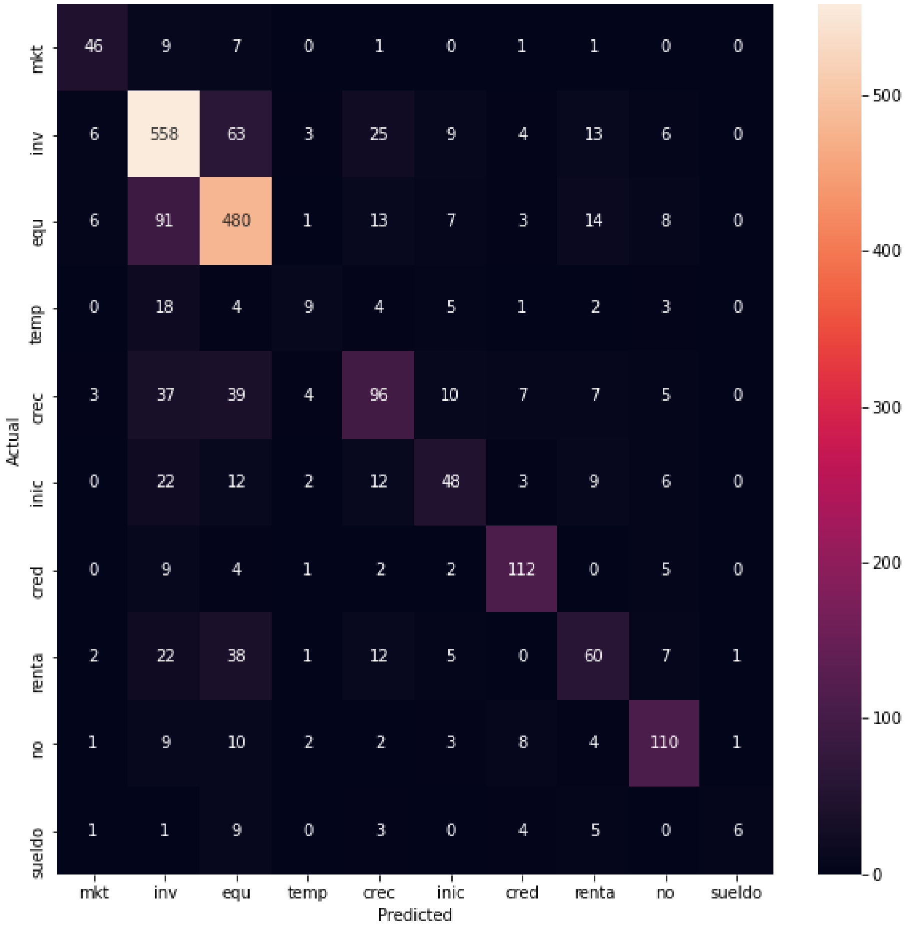


Figure 6: *F1 score of the different ML classifiers during cross validation.*

(Elegir qué métrica uso para comparar)

The best performers were Logistic Regression and Linear SVC, with almost 70% accuracy. It is worth noting that Logistic Regression failed to converge. However the results are stable between iterations. Given that the text was processed with tf-idf, it is unlikely that scaling would help fix this issue. However dimensionality reduction techniques like Latent Semantic Analysis (LSA) might be worth a try. Increasing *n_iter*, *tol* or changing the solver seem more promising in the first place. (Cambiar el solver de lbfgs a newton-cg lo solucionó)

We further evaluate the best performing model (Logistic Regression), by looking at its confusion matrix (figure 7).

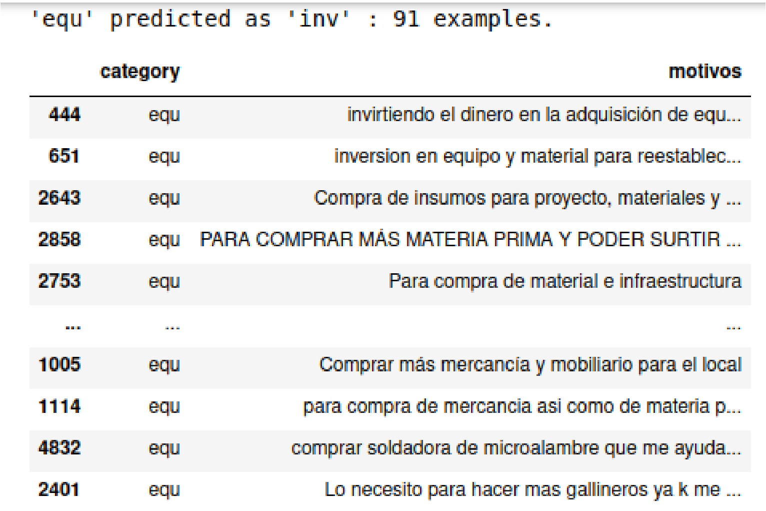


Actual \ Predicted	mkt	inv	equ	temp	crec	inic	cred	renta	no	sueldo
mkt	46	9	7	0	1	0	1	1	0	0
inv	6	558	63	3	25	9	4	13	6	0
equ	6	91	480	1	13	7	3	14	8	0
temp	0	18	4	9	4	5	1	2	3	0
crec	3	37	39	4	96	10	7	7	5	0
inic	0	22	12	2	12	48	3	9	6	0
cred	0	9	4	1	2	2	112	0	5	0
renta	2	22	38	1	12	5	0	60	7	1
no	1	9	10	2	2	3	8	4	110	1
sueldo	1	1	9	0	3	0	4	5	0	6

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
...	...	...
1005	equ	Comprar más mercancia y mobiliario para el local
1114	equ	para compra de mercancia asi como de materia p...
4832	equ	comprar soldadora de microalambre que me ayuda...
2401	equ	Lo necesito para hacer mas gallineros ya k me ...

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.

An algorithm may not perform well right out of the box but will with the right hyperparameters

Even though the sets are stratified, it appears that some labels are not present in the predictions. Generally this is due to imbalance in the dataset. However, stratification was performed when splitting and during cross validation. Further investigation is required (Pero ahora no da el tiempo)

5.2 Deep Learning

Usar F1, y actualizar las figuras

Figures 9 and 10 show the accuracy and loss curves respectively. On figure 9 we can see strong overfitting (A qué se debe, cómo solucionarlo). Figure 10 shows that loss is going up on the test set. This is another indicator that the model is overfitting. The following steps are reccomended:

- Reduce model capacity.
- Add regularization.
- Check that the training and test splits are statistically equivalent.

Probar estas cosas!

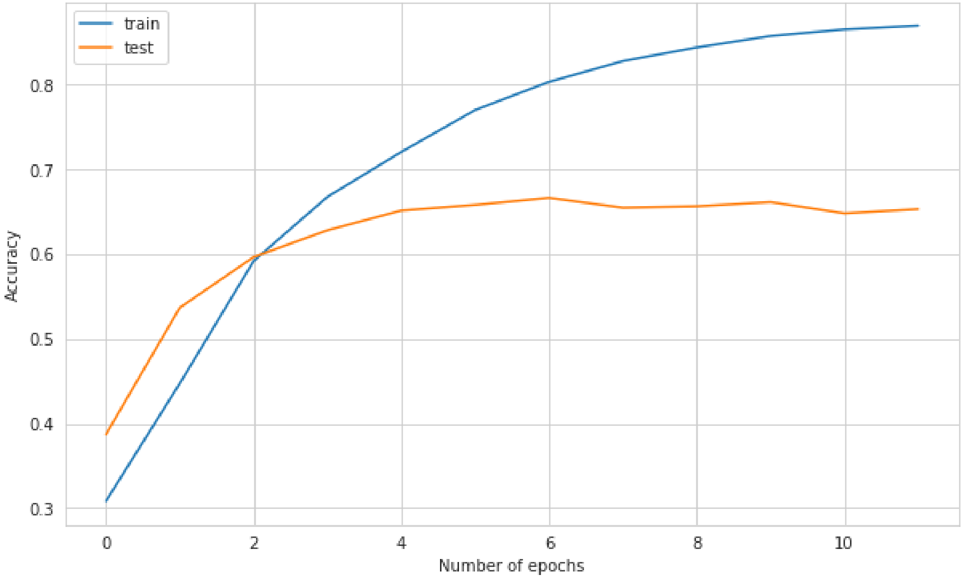


Figure 9: *Accuracy curve.*

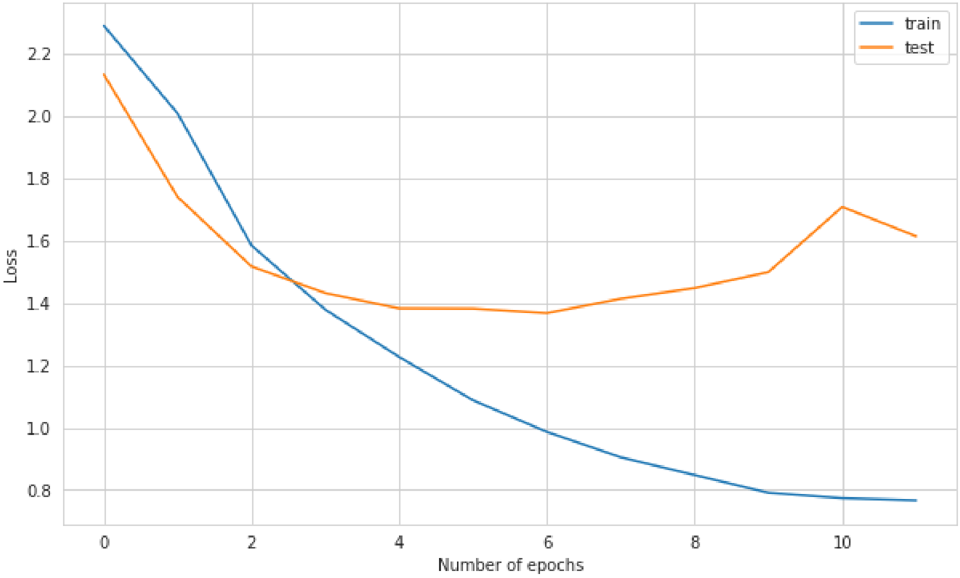


Figure 10: *Loss curve.*

6 Conclusion

Resumen del performance de los distintos modelos, causas (feature extraction/language model elegido) y cómo solucionarlo.

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.