# Sentiment Classification using Cross-Lingual Embeddings

**Thomas Ehling A20432671**
Chandana Ravindra Prasad, Sandeep Fnu, Inigo Alonso Gago

*CS 522: Advanced Data Mining, IIT Chicago*
Submitted 18 April 2019

## Abstract

We consider the problem of performing binary sentiment analysis on Spanish movie reviews using the MUSE cross-lingual embeddings and an LSTM network trained on an English dataset. Because data is scarce in foreign languages, the most common method remained using direct translation. We use datasets from 2 different sources and the more generalized 50_00 embeddings. Our model scored an accuracy of 72.97% on the Spanish reviews, against 87.56% on English reviews. This result shows that is possible to obtain a correct accuracy on reviews from a different source and language without translation. It brings a lot of new opportunities in a field needing huge amounts of data.

## 1 Introduction

Nowadays, the large majority of documents and services are available online. The main opportunity brought by globalization is the capacity to reach anyone in the world, but a direct consequence is the need to manage foreign languages.

Wherever humans are involved, languages become important. It is important for many aspects involving the common good, like health services, especially with the current state of worldwide immigration. In the industry, companies want to offer services to foreign customers (ex: recommendation on online retailers). For researchers and scientists, foreign datasets could bring valuable information but remain unusable.

Our project focus on the main challenge for data classification of foreign languages: data scarcity. In order to achieve high accuracy, a large amount of documents is required. Our context is the same as in the case of the problem we want to solve; how to classify a small dataset with a large dataset of documents in two different languages. This ultimately requires to process the datasets of the original and foreign languages, so we can operate on both in a similar way.

The actual solutions rely mainly on translating the smaller dataset into the same language as the other one. The translations can be manual or automatic. Manual translation introduces bias, as the understanding of the sentences and the resulting translation may change from one person to another. Furthermore, it is really expensive, difficult for large datasets and impossible for real-time operations. Automatic translation is one of the most active subjects of research, leveraging the power of new machine learning models. It is mainly used as a temporary solution, as there is still a low quality for the translations, due standard NLP (Natural Language Processing) and NLU (Natural Language Understanding) challenges (eg. Meaning, context, diversity, sarcasm, …).

In this paper, we analyze the efficiency of using the MUSE cross-lingual embeddings to classify documents with different datasets of different languages for training and testing. The challenging aspects of this problem are the lack of parallel corpus, no direct translations, and the use of pre-trained embeddings, that are known not to generalize well. Apart from presenting our results obtained via our classifier, we also analyze the problem to gain a better understanding of how difficult it is.

## 2 Previous Work

The first popular pre-trained embeddings have been the *GloVe: Global Vectors for Word Representation* (2014). That is the first time pre-trained embeddings generalized so well and could actually be used for classification.

More recently (April 3th 2019), the Amazon Alexa team released a paper: *Cross-lingual transfer learning for spoken language understanding.* They used cross-lingual embedding along six different machine learning architectures to show that cross-lingual transfer learning can reduce the data requirements for bootstrapping a spoken language understanding (SLU) system in a new language by 50%.

## 3 A closer look at our project

The two languages we selected are English for the training set and Spanish for the testing set. The motivation behind this selection is the current needs for this translation in the United States, and the fluency in both languages among our team.

The sentiment classification will be on movie reviews, due to the easy access to sample code, English datasets, and/or movie review websites to collect on our data. Also, movie reviews do not require too specific lexicon. The Classifier implements the MUSE embeddings and will be trained on the English dataset and tested on the Spanish one. We will use as a reference for our example, the accuracies obtained by classifying English reviews from a testing set and from a direct translation of the Spanish reviews.

## 4 MUSE Embeddings

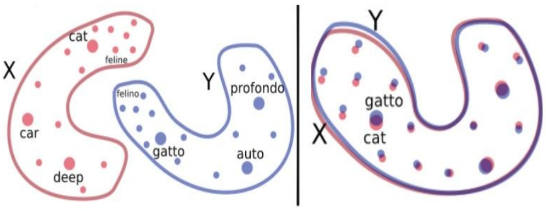


*Figure 1: Toy illustration of the MUSE embeddings before and after alignment.*

Embeddings are a multi-dimensional vector representation of words. They are widely used with Deep learning models for text classification, to quantify the relation between the words. .The MUSE embeddings are the pre-trained fasttext facebook words embeddings that have been aligned in the same dimensions. As we can see in Figure 1, once aligned, similar words from different languages will be close to each other.

The MUSE embeddings have been originally trained on the Wikipedia database, giving the embeddings a really neutral base. It is fitting for Sentiment classification because the words inside the vocabulary are very diversified, and the vectors have been trained without any specificities. Furthermore, movie reviews include few domain-specific words, mainly adjectives used in our everyday lives, and therefore, likely to be in the embeddings.

We will use during this project two sets of MUSE embeddings, one in English and one in Spanish. Each set is composed of 50 000 embeddings and 300 dimensions. For the preprocessing, punctuations and stop words are included. The case is ignored.

## 5 Datasets and data preprocessing

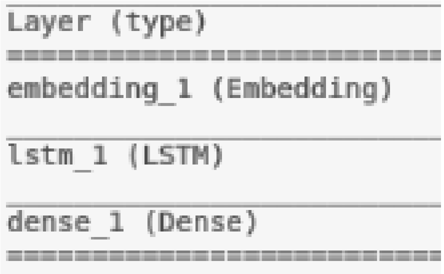
The English dataset we selected is the “IMDB” review dataset. and the “CorpusCine” for the Spanish one. To match the “IMDB” dataset, the “CorpusCine” files have been formatted to UTF-8 and converted from singular XML files to a global CSV file. The “IMDB” dataset is labeled with positive and negative sentiments (binary) where the “CorpusCine” reviews are labeled with the star ratings out of five. We set the threshold to 3, labeling “negative” for less, “positive” for more, and leaving out exact 3-star rating. It is important to keep balanced datasets so we selected 25_000/25_000 positive/negative reviews for “IMDB” and 1_130/1_145 positive/negative reviews for “CorpusCine”. The “IMDB” dataset is split with 40_000/10_000/10_000 for the training, validation and testing set.

Even if the “IMDB” dataset is available through the Keras library, we use the original raw data to have full control of the preprocessing steps. Our preprocessing is really specific to fit two challenges: filling the requirements of sentiment analysis and matching the words from the embeddings vocabulary. First of all, the Spanish special characters have been replaced the same way they are in the embeddings vocabulary. Then, we choose to format the reviews to lower case, to match the vocabulary and because it does not matter for positive and negative sentiments. No stop words have been removed, to protect the expression of negative feelings. Stemming or modifying the words directly by appending prefixes or suffixes will cause the words not to match the embeddings vocabulary. Finally, a review length analysis on the “IMDB” dataset show a mean of 240, an std of 179 and 75% of the reviews with less than 293 words. We used padding with zeros to apply a default size of 300 for all reviews.

## 6 Word Encoding

To perform the word encoding, we rely on the 4 vocabularies matching words to index: one by datasets and one by set of embeddings. We created the word to index vocabulary for “IMDB” and “CorpusCine” datasets according to word frequency, so we can still have the option to select only the most popular word later. Next, we updated the English embeddings vocabulary index to match the indexes from the “IMDB” one, keeping only the common words. Then, For each common word of the “CorpusCine” and Spanish MUSE embeddings vocabulary, we found the closest English MUSE embeddings and updated the indexing accordingly. With this final setup, we encoded all the reviews and created the embedding matrix.

## 7 ML model



```
Layer (type)
============================
embedding_1 (Embedding)

lstm_1 (LSTM)

dense_1 (Dense)
============================
```

*Figure 2: Illustration of the network architecture*

The Classifier here is a tool to highlight the efficiency of the MUSE embeddings. With this objective in mind, we selected a classic, simple LSTM (Long Short Term Memory) network, already well-tested for sentiment classification and suitable for pre-trained embeddings implementation.

RNN (Recurrent Neural Networks) are networks with loops in them, allowing the information to persist. LSTM networks are a special kind of RNN, capable of learning long-term dependencies. In the context of NLU, LSTM considers the order of the words in a sentence and allows the model to better understand the context or meaning of the sentences.

The finale model architecture is represented in Figure 2. The first layer contains the embeddings. According to the test, we either set the number of features and let the model learn his own embeddings and set the weights to the MUSE embedding matrix (cf. Word Encoding), and freeze the layer. The LSTM does not have any hyper-parameters but relies heavily on the text-preprocessing. The metric used for our project is the accuracy.

| MUSE embeddings | Test set | Accuracy |
|---|---|---|
| NO | IMDB | 86.94% |
| YES | IMDB | 87.56% |
| YES | Corpus Cine Translated | 87.26% |
| YES | Corpus Cine | 72.97% |

*Figure 3: Summary of the experiments results for 10 epochs*

## 8 Evaluation

See Figure 3 for a summary of all the experiments accuracy score. we ran all the models for 10 epochs, and 300 features.

The main difficulty we encountered during experiments is the computation time, the text preprocessing and training of the models take hours. This made us focus on the most essential features we wanted to test and implements, as each run matters.

The classifier performs slightly better with the MUSE embeddings that without. This is counter-intuitive as embeddings tend to be better when not generalized, especially for sentiment analysis. This could be explained by the relatively small size of the dataset (50_000) or/and also by the too high number of features we selected (300). We actually noticed that we overfit really quickly without the muse (the graph is available in the appendix).

Our final accuracy is strictly lower than the other ones. That could be explained by a few challenges we encountered along with the project. The biggest unexpected result is the out of vocabulary percentage: 80% for English 67.4% for the Spanish dataset. Even if it seems to be high, it ends up being ok for the English, as the final number of the common words is 35_741, which means we are using 71.5% of the embeddings vocabulary. But for the Spanish, there are only 20_385 common words, that represents 40% of the Spanish embeddings vocabulary. We can correlate this result with the fact that all the specific Spanish characters ( e.g.: ñ ) have been removed from the words. The code to convert these is implemented but the data of the original dataset is corrupted. We chose to stick with this dataset because it is labeled, balanced, large enough and is specific to movie reviews. It could be possible to match the words with a missing character to the corresponding one in the embeddings vocabulary, but the computation time it will take is too long to be worth it. We think that the high accuracy may be due to the fact that out of bag words belong to the less frequent ones, being too specific.

The translated Spanish accuracy is here to compare the performance of our method with the current process in the industry. We translated the review with the Google translate API, one of the best translation tool available. It can easily translate the words with missing characters, by understanding the context. So it is expected to have higher accuracy. However, we would like to highlight 2 points. First, the goal is here to prove that using the cross-lingual embeddings may be good enough for some applications to use it instead of doing a direct translation first. Finally, we are here using Spanish, so the translation is relatively easy, but the power of the embeddings is to be applicable to distant languages like Chinese.

## 9 Result Analysis

*9.1 Score Evaluation*

Humans disagree among themselves about the sentiment of an online post 10% to 30% of the times. This means that depending on sarcasm and ambiguity of the post the sentiment accuracy should be anywhere between 70% and 90%. Considering all the constraints and challenges of our project, 72.97% is an acceptable accuracy. The difference with English reviews remains 15%, it is relatively important considering that the English reviews also come from the IMDB dataset.

*9.2 Past errors and what we learned from it*

This part is dedicated to the challenges we faced and corrected along with the projects. First, for the IMDB dataset, we designed the preprocessing steps implementing the Keras library function. But it was clear, by looking at the sentences, that the preprocessing was not the same. To prevent any bias and offer a fair comparison, we decided to create our own preprocessing pipeline.

Then, during the preprocessing, we wanted to remove as many stop words as possible to reduce the size of the training set. However, some stopword remains important in sentiment analysis, especially when using an LSTM. We first wanted to remove all stop words except the one for the negation but we were not able to apply the same process to the spanish reviews efficiently. To preserve the same preprocessing, we kept the stop words in both datasets.

Finally, In order to save some processing time, we match the Spanish embeddings to the closest English ones relying only on the 100 first embeddings. We selected this number as it was the threshold where a list of 20 Spanish words was matched to their closest translation. Once the code finished, we updated the vocabulary using the 300 embeddings and the accuracy improved from 72.44 to 72.97.

*9.3 Actual errors and possible solutions*

| | Predicted Pos | Predicted Neg |
|---|---|---|
| Actual Pos | TP: 1014 | FP:131 |
| Actual Neg | FN: 484 | TN: 646 |

*Figure 4: Confusion matrix*

Our confusion matrix is visible in figure 4. We also calculated the F1 score to see if the recall was important. The final F1 score is 76.73, approximately the same as the accuracy. We concluded that the recall was great and did not have much impact on the analysis, so we kept the accuracy scores.

The main criteria that affect our accuracy is the small size of the final Spanish vocabulary. One main reason for it is the size of the dataset. However, this is one of the challenges our project aims to tackle, so taking a larger dataset cannot be the right answer, even if it will probably increase the accuracy. Then, using the embeddings is limiting ourselves to a lexicon, the one provided by the embeddings vocabulary. So, no matter how great the model is, our vocabulary will still have a maximum size of 50_00, and will only contain general words. This is also a requirement of our project, but we wanted to precise that our final solution may not be suitable for more specific sentiment analysis.

| | | | |
|---|---|---|---|
| **TN** | 51.08 | **FN** | 48.40 |
| **TP** | 50.53 | **FP** | 49.55 |

*Figure 5: Average number of unknown characters by review.*

As stated before, most of our Spanish words are out of vocabulary. Because of the poor quality of the original dataset, all Spanish special characters have been deleted from our data. However, we can still count how many there are by review, and analyze the actual impact on our model. We wanted to see if reviews with more unknown characters have been more heavily misclassified. The result of our analysis is represented in Figure 5. We learned 2 valuable pieces of information from it. First, the number of unknown character did not influence our model, as there is the same ratio in each category. That's great and explain the high value of our final accuracy. However, the total average of unknown character by review is higher than we thought, 50 by review! It means that we dropped an average of 50 words by review and is most likely the reason for our out of vocabulary ratio. The solution should be to try another processing, maybe manual, of the original dataset to preserve the accent, or select another dataset of better quality. The solutions we already implemented to reduce the number of out-of-vocabulary words, is to remove stemming and lemmatization from the preprocessing steps, as it would have modified the structure of the words.

Keeping a balanced dataset is really important, even more in data analysis. The original IMDB dataset is balanced with 25_000/25_000 positive/negative reviews, but we noticed that we shuffled the data randomly before splitting it into the testing/validation/training sets. Maybe one of the set ended up being unbalanced and can affect our final results. But after verification, the sets are perfectly balanced with a repartition positive/negative review of 20_018/19_882, 2_471/2_529 and 2_511/2_489 respectively. This was indeed not important for binary classification, the probability of belonging to a class after shuffling being 50%. However, it is an important aspect to consider if we would generalize it to more classes.

| | Percentage of truncated reviews |
|---|---|
| TP+TN | 73.40 % |
| FP+FN | 26.59 % |

*Figure 6: Percentages of truncated reviews*

The last important decision we made that impacted the accuracy is the review length after preprocessing. Knowing that more than 75% of the English reviews were shorter than 293 words, we selected a size of 300. However, the Spanish reviews are much longer, with a mean of 437. Further analysis revealed than 69.58% of the Spanish reviews were truncated. Furthermore, the percentages of these truncated reviews among the classified and misclassified reviews are displayed in Figure 6. We can see that the size does influence the final accuracy, as the classifier tends to classifier more accurately the longer ones. An interesting experiment would be to increase the length and see if the accuracy does increase.

Most of these errors are directly related to the Spanish dataset we used. Finding a labeled Spanish movie review dataset was not easy, and we are glad that we have found this one. However, the best situation would have been to be able to use two datasets from the same source, so the people among the community, the vocabulary words, and the review length are similar. (e.g. Netflix reviews in English and Spanish).

Finally, some Spanish sentences extracted from rightly classified and misclassified reviews are available at the end of the appendix.

## 10 Discussion

Multilingual techniques, including Cross-lingual word embeddings, are an active area of research, full of opportunities. Only a few weeks ago, the Amazon Alexa team released a paper on multilingual text classification with word embeddings, character embedding, and several deep neural networks architectures.

The specificity of this project was to show the efficiency of the MUSE embeddings, using only the 50_000 most general ones and training and testing on two datasets from different language but also different sources. Even with all these challenges, we still ended up with a high accuracy of 73%, 15% different from the testing on English data. Our result proves that we can use the MUSE embeddings to classify an unseen dataset in a new language.

This is really exciting. These MUSE Embeddings have been made available by Facebook, so we can predict a future where the machine Learning model can be developed and trained by large companies like Amazon, on their billions of reviews, and we could use these model to classify any document in any language! This is an even more amazing breakthrough as we need new translation technique with the world globalization and immigration status, and it can bring new discoveries in languages that were not accurately useable due to the scarcity of the data.

## References

Github for the project :
https://github.com/ThomasEhling/Cross-Lingual-Embedding

A. Conneau*, G. Lample*, L. Denoyer, MA. Ranzato, H. Jégou. 2018 *“Word Translation without Parallel Data”*

Jeffrey Pennington, Richard Socher, Christopher D. Manning. 2014 *“GloVe: Global Vectors for Word Representation”*

Quynh Ngoc Thi Do, Judith Gaspers. 2019 “*Cross-lingual transfer learning for spoken language understanding.”*

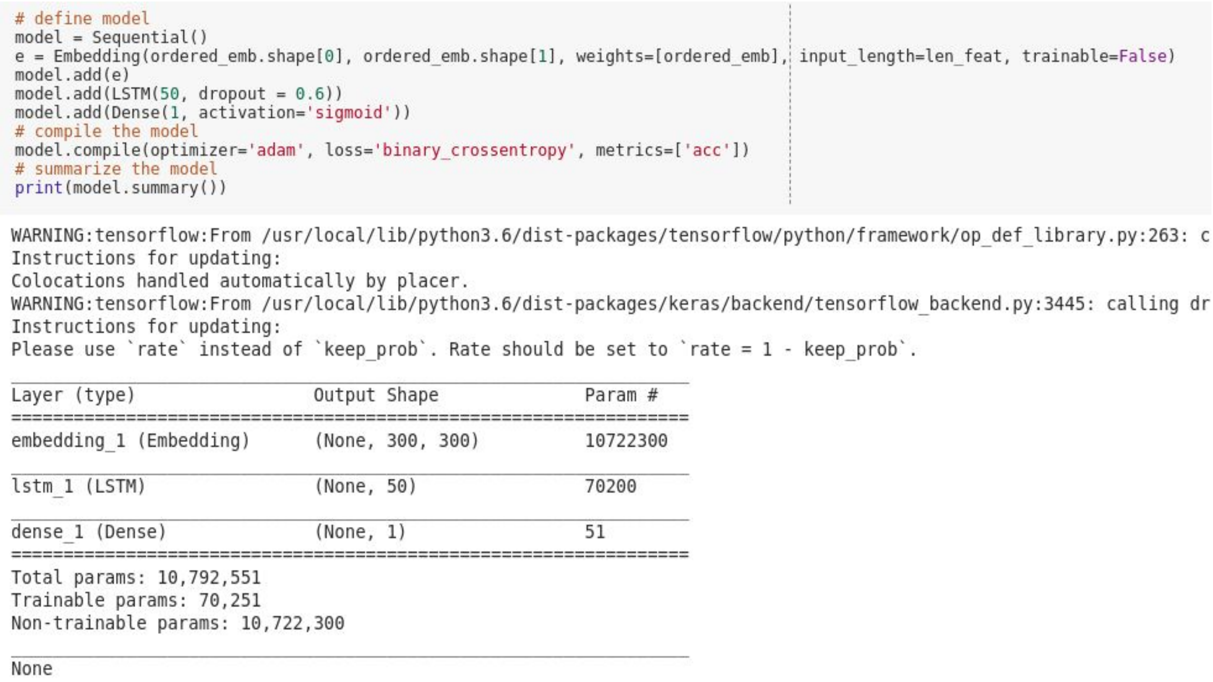
- MUSE embeddings GitHub page :
https://github.com/facebookresearch/MUSE

- IMDB review dataset :
https://ai.stanford.edu/~amaas/data/sentiment/

- Corpus Cine Dataset :
http://www.lsi.us.es/~fermin/index.php/Datasets

# APPENDIX

Here is a screen shot of the LSTM definition :

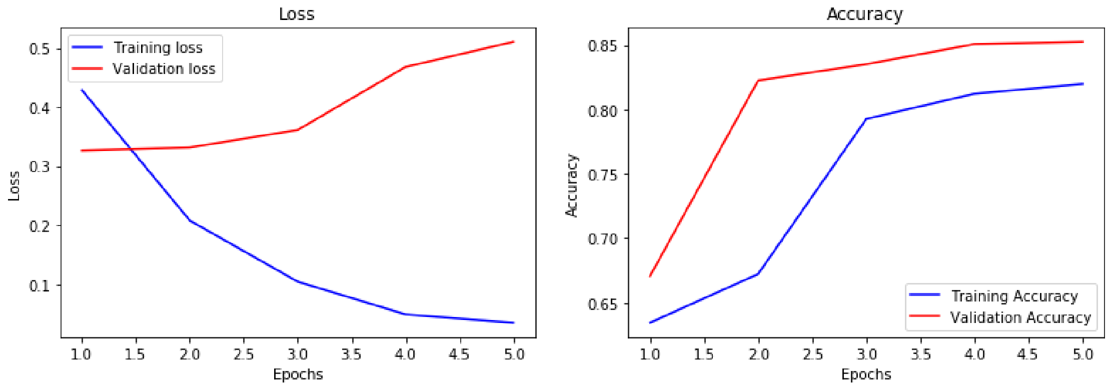


```
# define model
model = Sequential()
e = Embedding(ordered_emb.shape[0], ordered_emb.shape[1], weights=[ordered_emb], input_length=len_feat, trainable=False)
model.add(e)
model.add(LSTM(50, dropout = 0.6))
model.add(Dense(1, activation='sigmoid'))
# compile the model
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['acc'])
# summarize the model
print(model.summary())
```

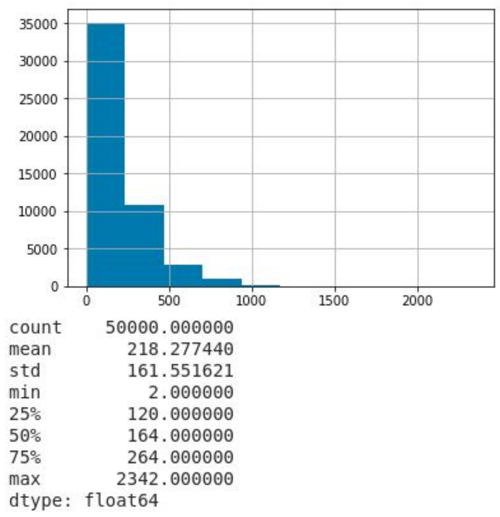
```
WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/tensorflow/python/framework/op_def_library.py:263: c
Instructions for updating:
Colocations handled automatically by placer.
WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/keras/backend/tensorflow_backend.py:3445: calling dr
Instructions for updating:
Please use `rate` instead of `keep_prob`. Rate should be set to `rate = 1 - keep_prob`.
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
embedding_1 (Embedding)      (None, 300, 300)          10722300  
_________________________________________________________________
lstm_1 (LSTM)                (None, 50)                70200     
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 51        
=================================================================
Total params: 10,792,551
Trainable params: 70,251
Non-trainable params: 10,722,300
_________________________________________________________________
None
```

All the cleaned code files, datasets, logs and plot can be find with this link : https://drive.google.com/drive/folders/1Q3Ifz1gtHVFKIUb7kCMIHQ7SoyZAEPJN?usp=sharing

For the most important experiments, we keep the plots with accuracy and loss by epochs. For example here are the plots for the model with the muse embeddings, and 5 epochs :

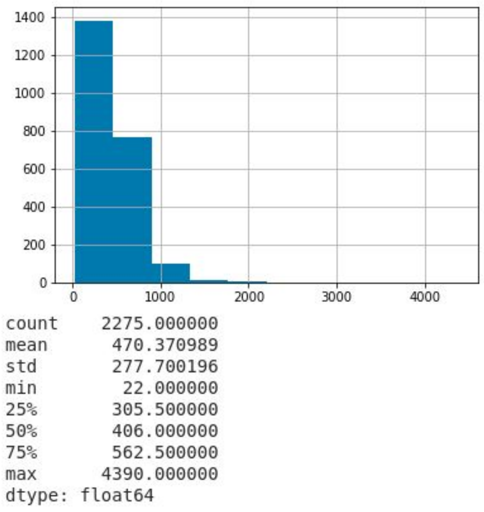


Here also the histograms of the reviews length :



```
count    50000.000000
mean       218.277440
std        161.551621
min          2.000000
25%        120.000000
50%        164.000000
75%        264.000000
max       2342.000000
dtype: float64
```

English dataset



```
count    2275.000000
mean      470.370989
std       277.700196
min        22.000000
25%       305.500000
50%       406.000000
75%       562.500000
max      4390.000000
dtype: float64
```

Spanish Dataset

Finally here is sample reviews, misclassified and not, with their translations :

| **Category** | **# words** | **Review** |
|---|---|---|
| FN original review | 1430 | Relata las inquietudes del hombre contempor�neo, con abundantes met�foras y momentos inolvidables La vida de Don, un Don Juan pasado en a�os, se ha convertido en una especie de estancamiento[...] El bueno de Bill es �nico en hacer gala de su inexpresividad |
| Translation : | | It narrates the worries of a contemporary man, abundant metaphors and unforgettable moments The life of a Don, a Don Jonh too old, has become stalled[...] The good Bill is unique showing off his inexpressiveness. |
| FP original review | 252 | bodrio bodrio bodrio bodrio bodrio bodrio bodrio bodrio bodrio siete fueron los guionistas reclutados para parir el engendro una con cabeza habr�a bastado […] se han embarcado en la peregrina idea de hacer un film novedoso[…] adi�s placeres sencillos |
| Translation : | | Shit shit shit shit shit shit shit shit what a shit the script writers were recruited to give birth to a monster one head would've been enough [...] they couldn't help making something original [...] good by to easy pleasures. |
| TP original review | 531 | Una historia que nos hace reflexionar sobre las casualidades , sobre la mala suerte , sobre c�mo te puede cambiar la vida en 1noche[..] |
| Translation : | | A story that makes us think about chance,about bad luck, about how life can change overnight[...] |
| TN original review | 257 | Una pel�cula que huye de la l�nea habitual en el cine espa�ol, pero que lo hace recurriendo a un argumento repetitivo [..] merece dedicarle un rato, pero no en el cine, donde no creo que vuelva a proyectarse, sino en el sill�n de casa  y con un gin-tonic y un paquete de pipas para pasar el rato. |
| Translation : | | A movie that differs from mainstream Spanish cinema, but it resorts to a plot too repetitive[...] it is worth some of your time, but not in the where I don't think it would be projected again, only sitting in your house with a gin-tonic and bag of seeds to have some fun. |