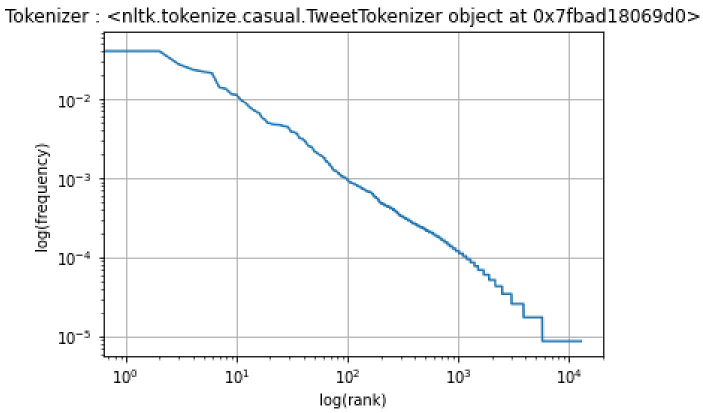
Sentiment Analysis : predicting the impact of a financial newspaper title

Jérémy Berloty, Amaury Dreux

Description of the dataset

Our dataset consists of 4845 titles of financial newspapers. They were anotated by experts in three different classes : ”neutral”, ”negative” and ”positive”. We transform them in numerical variable for further processing. To describe the dataset, we will check the Zipf’s law, the words that occur the most and finally the words that are the most associated with a negative or positive review. To do so we tokenise the sentences with the tweet tokeniser technique. In the following figure we clearly see that the occurences are inversly proportional to the ranking of the words (in terms of occurencies). Therefore our dataset follows the Zipf’s law. After only



keeping the nouns and adjectives, we can plot the words according their occurences with a wordcloud, as shown below. Some of the words that occur the most are ”EUR”, ”company”, ”Sales”, ”Profit”, ”Year”, ”Period” ...

Finally the figure below shows the words most associated with negative reviews (for words with at least 10 occurences). The index is the mean of the appearences in the different kinds of reviews (*(appearences in positive reviews)+(appearences in negative reviews)/(total appearences)*). Thus it is between -1 (for the most negative ones) and 1 (for the most positive ones. The most negative words are for instance, ”warning”, ”temporarily”, ”decreased”, and more surprisingly ”jobs”, ”workers”, ”staff” or ”employment”. Marx could have been interested by this detail. The most positive words are for instance ”narrowed”, ”grew”, ”efficient”, ”awarded” ...

On average the words are more associated with positive reviews (0.2 approximately), which can be explained by the fact the data is unbalanced : there are **604** negative reviews, **1363** positive reviews and **2872** neutral ones. The unbalanced data could also be a problem for the prediction algorithm developed after.

Prediction and leads of improvements

	word	occurences	freq	type	scale
11343	warning	11	0.000095	VBG	-1.000000
4690	temporarily	14	0.000121	RB	-0.928571
5133	lay	11	0.000095	VBD	-0.909091
535	decreased	79	0.000683	JJ	-0.835443
858	fell	55	0.000476	VBD	-0.800000
3581	dropped	15	0.000130	VBD	-0.800000
3346	lower	39	0.000337	JJR	-0.666667
8644	lay-offs	17	0.000147	JJ	-0.647059
2779	fall	14	0.000121	NN	-0.642857
1098	declined	16	0.000138	VBD	-0.625000
7851	cut	22	0.000190	NN	-0.590909
3586	jobs	17	0.000147	NNS	-0.588235
4578	Scanfil	26	0.000225	NNP	-0.576923
2731	temporary	18	0.000156	JJ	-0.555556
50	workers	13	0.000112	NNS	-0.538462
2829	staff	32	0.000277	NN	-0.500000
2601	ADPnews	12	0.000104	NNP	-0.500000

For the prediction model we choose to implement an LSTM. The embedding of words is done from a pretrained vectors from Fast Text. The LSTM takes as input a sentence of embedded words (of dimension 300) and outputs a vector of size 268*(batch size). Then a fully connected layer gives 3 outputs (the three classes), which is fed into a softmax function which gives the final output. The final output is three probabilities, for each class, and the predicted class of the input will be the highest probability. To fit the model we use a cross entropy function, and we use an Adam Gradient Descent with a learning rate of 0.001. To avoid overfitting we add a dropout function. The dataset is split in a train dataset (80% of the dataset), and an eval and test dataset (10% each).

Unfortunately the model does not give good results. The loss is not reducing (stuck around 0.9), and it only predicts "neutral". One of the reason of this poor result we thought of was the unbalanced data as mentionned before. The leads to improve our results were theefore to do prediction only on the negative and positive reviews, or to undersample the positive review. Another limit could come come from the model : indeed recurrent neural networks struggled to get long term relationships and have vanishing gradient issues. A transformer model like BERT could prove itself more efficient.