Topic classification with news article headlines
Machine Learning for Natural Language Processing 2021

Vincent Burnand Galpin
ENSAE 3A
vincent.burnand.galpin@ensae.fr

Litti Esteban
ENSAE 3A
litti.esteban@ensae.fr

1 Problem Framing

Today, online articles are a dime a dozen: thanks to digitalization, continuous information has become the norm. We are inundated with information and news. Thanks to a classification model based on the BERT pretrained model, we want to automatically categorize articles into categories in order to better target the information according to the readers' interest profile. More precisely, in order to improve the efficiency of the algorithm, we want to see if the headline of the article is enough to categorize an article according to its category or if it is necessary to use the whole article to be able to categorize it properly. Our question will be : is a news article headline enough to determine the topic of an article?[1]

2 Experiments Protocol

We use news article database compiled by Kishan Yadav[2], composed of english short articles relative to indian and international news, from a webapp. After removing duplicates, our dataset contains 5111 short articles. For each, we have the headline, the article and the category. There are 7 categories (associated to a label between 1 to 7). Our articles are mainly in the categories entertainment, world, sports and technology, and less in the categories science and automobile.

On average, an article has a headline of 11 words, and the article itself is very short : on average 58 words, and 61 for the maximum, certainly due to a fixed limit imposed by the webapp.

After exploring the data, we construct a topic classification, first using only the title, and then using the title and the article, in order to see if the whole article significantly improves the prediction or if the headline is sufficient. We tokenize using bert-base-uncased and then we use BERT for sequence classification, with Hugging Face library. As an output, we have for each article the probabilities of belonging to each of the 7 categories. We tried 2 different classification decision methods : first, assigning the article to the category for which the probability was higher than 0.5. This was satisfying, but in some cases where the highest probability was under 0.5 (1.5 to 2.5%), the model assigned no predictive label. Looking closer to these examples, we saw that the first choice was often the right one. For example, there is an article with the title "Disney builds robot with life-like human gaze" in the technology category : the model attributed a 37% chance to the article to be technology and 23% to entertainment. The model was hesitant but it was right. We thus corrected the decision method by assigning the label which has the highest probability among the 7 labels.

We then conduct a quantitative evaluation, using the F1 score on the test dataset. We also conduct a detailed qualitative evaluation, looking to specific mistakes and the confusion matrix. Finally, we compare the results obtained between using just the title and using the title and the article, on the same train and test datasets, using quantitative and qualitative methods.

Category	Label	Number of articles
entertainment	1	1133
world	2	1118
sports	3	1011
technology	4	911
politics	5	570
science	6	239
automobile	7	129

[1] GitHub: https://github.com/VincentBurnandGalpin/NLPprojecttopicclassificationcttopicclassification, Colab: https://colab.research.google.com/drive/1oSSbN1yhlithxSev1jiRijIps_tvhYCN?usp=sharing

[2] https://www.kaggle.com/kishanyadav/inshort-news

3 Results

Using only the headlines, we get a F1 accuracy of 0.908. The more frequent the label is, the better the models F1 score is - the model has the most difficulty in predicting the categories automobile and science. We qualitatively evaluate our results, looking at specific examples. We see that in many misclassified articles, the model was not very sure about the classification (the predicted label scores are all under 0.5). For instance for the headline "dior features depp as face of cologne following 'wife beater' ruling", the model classifies in technology with a probability very low of 0.26. Also, we can see that in many cases, the second highest predicted value is the good one. For instance, the headline "meteor shower streaks across the night sky in UK on christmas" is classified as world (probability 0.87), but the second highest probability (0.14) is for science which is the good label.

We also look at the confusion matrix to see which categories are confused. For instance, sport headlines (label 3) are sometimes classified as entertainment (1), which make sense. Also, in the science category (6), 7 headlines out of 36 are in the world category (2). When we look closer to these 7 cases, we understand that this happened because the name of a region was given in the title (US, Amazon, British, UK), what lead our model to classify as world, while the true category was science.

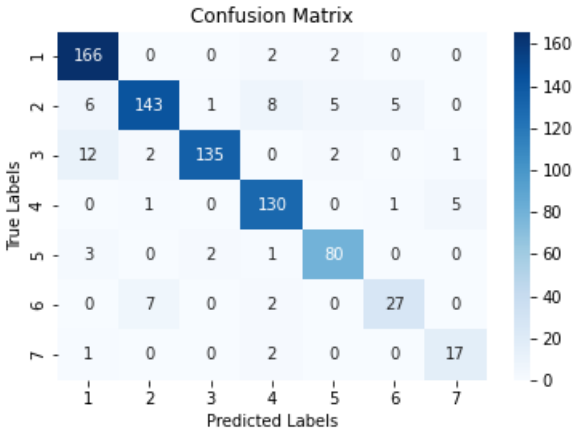

Figure 1: Confusion matrix, using only the headlines

We then rerun the model on the same train and test datasets, using the title and the article, after reinitializing the BERT model in order to avoid overfitting. We get a F1 accuracy of 0.932, to compare to the 0.908 obtained using only the headlines. The accuracy is improved, as could be expected because we provide more content. However, the improvement is only of a few points. We then conduct a qualitative comparison between the results of the model using only the title, and using the title and the article. There are 38 cases where the predicted label by only titles is false and the predicted label by the whole article is correct (and 19 cases where it is the opposite). When we look at examples where the article enables the model to perform better, we realize that the article contains key words that are not in the title. For instance, the article with the headline "got covid - 19 after being locked down with republicans during riots : jayapal", we might think that the model only using the headline sees the world "republicans" and classifies as politics. However, in the article itself, words such as "US", "Capitol", "Indian" allow our model to understand that the category is world (the right label).

4 Discussion/Conclusion

The difference between the F1 accuracy of the model with the title and with the title and the text is very small, while the title contains on average 6 times less words. We can reasonably assume that the headline already contains all the most important keywords of the article, because the purpose of a headline is to catch the eye so that the text is read by the reader. We can thus conclude that the headline is enough to classify the topic of an article, because reading the whole article only improves the performance by 2.4 points, and it implies reading 6 times as many words. A more efficient solution would be to use only the headline, and if the predicted probabilities are all under a threshold (say 0.6), meaning the model is not very sure, the article is read.

In order to improve our model, we could have increased the number of epochs. Indeed, with 3 epochs, it is possible that a bad start of the model is not caught up. However, we made this choice to avoid having a heavy train phase, and because the performance was already satisfactory. Furthermore, we could improve the model by including a name entity recognition process. This would have allowed us to quickly identify proper names associated with certain themes, such as Elon Musk for the technology or automobile category. This is a future area of research that could complement our already encouraging results.