# TEXT CLASSIFICATION

NLP PROJECT

Author: Javier Gallego Gutiérrez
Course: Intelligent Systems
Degree: Master's Programme in Data Science
Date: January 30, 2022

## 1 Introduction

Natural Language Processing (NLP) is a Computer Science branch that works with human language in order to perform useful tasks, such as human-machine communication or simply processing text or speech [1]. It has experienced a revolution in recent years due to deep neural network machine learning methods. Along this practical application we will try some NLP techniques with the aim of doing a particular task.

## 2 Problem description

The main goal for us will be to perform text classification over a data set with media headlines trying to identify those that are considered clickbait. The data set used [2] is available at `https://github.com/bhargaviparanjape/clickbait`. It has 32000 different headlines where 16001 are non-clickbaits and 15999 are clickbaits.

## 3 Experiments

For the experiments performed we used Flair [3], a Python NLP package. The code developed can be seen at `https://github.com/javiegal/text-classification`. The process followed is depicted in the following lines and is based on two tutorials: `https://towardsdatascience.com/text-classification-with-state-of-the-art-nlp-library-flair-b541d7add21f` and `https://github.com/flairNLP/flair/blob/master/resources/docs/TUTORIAL_7_TRAINING_A_MODEL.md`.

First of all, we have to join the clickbait and non-clickbait files in one data set. Then, we will split it in three different sets, in the usual way data sets are split for Machine Learning tasks: training, validating and testing. We will use 5% of the data set for training (1600 headlines), 5% for validating (1600 headlines) and 90% for testing (28800 headlines). Additionally, these three sets have to be stored in FastText format.

After that, we will create the corpus. For that purpose, we will use the Flair class called `ClassificationCorpus`. Then, we will use document-level embeddings for our headlines because they are better than word embeddings for this type of task. These embeddings give you one embedding for an entire text. Among all the document embeddings Flair provides, `TransformerDocumentEmbeddings` is the chosen type. It uses pre-trained transformers and it is the recommended one for text classification by Flair developers. The model used in our case will be the standard BERT-Base uncased transformer [4] and we will fine-tune it with our data. At this point, the `TextClassifier` model can be created and trained with a `ModelTrainer`. We will do it with a learning rate of 0.001 and 3 as the maximum number of epochs. We limited the number of epochs because `TransformerDocumentEmbeddings` without a GPU take a really long time. Once the model is trained, we will get a summary with the performance measures obtained for the training set.

## 4 Results

Listing 1 show the results obtained for the training set with our model. These results look extremely good and we can see how 3 epochs were enough to get an accurate classifier. On the other hand, the time taken for training and testing the classifier was a bit long: around an hour and a half.

```
Results:
- F-score (micro) 0.9998
- F-score (macro) 0.9998
- Accuracy 0.9998

By class:
              precision    recall  f1-score   support

          cb     1.0000    0.9995    0.9998     14363
         not     0.9995    1.0000    0.9998     14437

   micro avg     0.9998    0.9998    0.9998     28800
   macro avg     0.9998    0.9998    0.9998     28800
weighted avg     0.9998    0.9998    0.9998     28800
 samples avg     0.9998    0.9998    0.9998     28800
```

**Listing 1.** Results obtained for the training set.

Unfortunately, we believe these results detect some structures frequently used in clickbait headlines that can be used in non-clickbait headlines. If we take a look at the data set, a lot of clickbait headlines from it contain numbers—e.g., "23 Delicious Ways To Layer Up For Fall". For that reason, we decided to check the model with a small set of particular headlines to get a better view about it. The nine headlines used were taken from different web sites. In the following lines we show them and our belief about each one being or not a clickbait headline:

1. "16 Men Died in New York City Jails Last Year. Who Were They?" Not a clickbait.

2. "Massive Price Change on this one!!! Almost NEW! Check it out – $259,900!!!" Clickbait.

3. "I CAN'T BELIEVE THIS HAPPENED... OMG!" Clickbait.

4. "Federal Judge Rejects Hate-Crime Plea Deals in Ahmaud Arbery Killing." Not a clickbait.

5. "We Know Why You're Single Based On Your Zodiac Sign." Clickbait.

6. "Covid-19: CDC warns against travel to 22 countries including Australia and Israel." Not a clickbait.

7. "Summer breaks: 20 of the best self-catering stays in the UK." Not really sure. It has a typical clickbait structure, but in the end it is just a ranking of touristic places. The body text gives you what you expect.

8. “Corona Centennial leads The Times’ top 25 prep basketball rankings.” Not a clickbait.
9. “18 Times When Stop Clickbait Has Saved the Day.” Clickbait.

Listing 2 shows the output obtained for the prediction of these headlines with our model. The label assigned to each sentence is underlined (**cb** if it is a clickbait and **not** if it is not). Headlines 2-6 and 9 are properly classified, but the other three require some comments. The first headline is wrongly classified. This headline corresponds to a journalistic report. It may be written with the aim of impacting the reader, but it is still a good summary of what the report offers. However, the value obtained (0.6604) was not as big as with the other true clickbaits. Headline 7 is classified as a clickbait (0.7899 value), but, as we said before, we are not really sure if it is so. Finally, headline 8 is correctly classified as a non-clickbait headline, but with a low value (0.543). These three examples let us think that rankings or enumerations are more susceptible of being classified as clickbait than other headlines. Of course, it is just a hypothesis and a detailed study should be done in order to accept or reject it.

```
Sentence: "16 Men Died in New York City Jails Last Year . Who Were They ?"   [-
     Tokens: 15  - Sentence-Labels: {'label': [cb (0.6604)]}]
Sentence: "Massive Price Change on this one !! ! Almost NEW ! Check it out - $
    259,900 !! !"   [- Tokens: 19  - Sentence-Labels: {'label': [cb (0.8031)]}]
Sentence: "I CAN'T BELIEVE THIS HAPPENED ... OMG !"   [- Tokens: 8  - Sentence-
    Labels: {'label': [cb (0.9804)]}]
Sentence: "Federal Judge Rejects Hate-Crime Plea Deals in Ahmaud Arbery Killing
    "   [- Tokens: 10  - Sentence-Labels: {'label': [not (0.9907)]}]
Sentence: "We Know Why You 're Single Based On Your Zodiac Sign"   [- Tokens:
    11  - Sentence-Labels: {'label': [cb (0.9084)]}]
Sentence: "Covid-19 : CDC warns against travel to 22 countries including
    Australia and Israel"   [- Tokens: 13  - Sentence-Labels: {'label': [not
    (0.892)]}]
Sentence: "Summer breaks : 20 of the best self-catering stays in the UK"   [-
    Tokens: 12  - Sentence-Labels: {'label': [cb (0.7899)]}]
Sentence: "Corona Centennial leads The Times' top 25 prep basketball rankings"
      [- Tokens: 10  - Sentence-Labels: {'label': [not (0.543)]}]
Sentence: "18 Times When Stop Clickbait Has Saved the Day"   [- Tokens: 9  -
    Sentence-Labels: {'label': [cb (0.9484)]}]
```

**Listing 2.** Results obtained for the training set.

## 5 Conclusion

Document embeddings and transformers allowed us to build a successful model that distinguishes between clickbait and non-clickbait headlines. However, as there is not a clear definition about what a clickbait is, this task may be more difficult and our results dependent of the data set used. One option for future work could be establishing the definition “a clickbait is the headline of a body text that does not contain the information the headline suggests” and try to make a classifier that somehow conforms to this definition using the headline and the body text. Other possible future task is to find a big enough set of non-clickbait headlines that contain numbers and check how they perform in our model.

Throughout this project, we were able to check some of the capabilities current NLP frameworks have. Anyway, we should pay attention, because it is evolving very quickly and the techniques used today may get outdated pretty soon!

## References

[1] D. Jurafsky and J. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, ser. Prentice Hall series in artificial intelligence. Pearson Prentice Hall, 2009, ISBN: 9780131873216.

[2] A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly, “Stop clickbait: Detecting and preventing clickbaits in online news media,” in *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, IEEE, 2016, pp. 9–16.

[3] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, “FLAIR: An easy-to-use framework for state-of-the-art NLP,” in *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 2019, pp. 54–59.

[4] I. Turc, M.-W. Chang, K. Lee, and K. Toutanova, “Well-read students learn better: On the importance of pre-training compact models,” *arXiv preprint arXiv:1908.08962v2*, 2019.