A Review of Text Classification Models from Bayesian to Transformers

Ema Ilic[1], Mercedes Garcia Martinez[1] and Marina Souto Pastor[1]

[1]*Pangeanic, Valencia, Spain*

Abstract

This paper is discussing a review of different text classification models, both the traditional ones, as well as the state-of-the-art models. Simple models under review were the Logistic Regression, naïve Bayes, k-Nearest Neighbors, C-Support Vector Classifier, Linear Support Vector Machine Classifier, and Random Forest. On the other hand, the state-of-the-art models used were classifiers that include pretrained embeddings layers, namely BERT or GPT-2. Results are compared among all of these classification models on two multiclass datasets, 'Text_types' and 'Digital', addressed later on in the paper. These datasets are internal to Pangeanic. The experiments were coded in Python 3.8. The codes have been executed with various quantities of data, on different servers, and on two different datasets. While BERT was tested both as a multiclass as well as a binary model, GPT-2 was used as a binary model on all the classes of a certain dataset. In this paper we showcase the most interesting and relevant results. The results show that for the datasets on hand, BERT and GPT-2 models perform the best, though the BERT model outperforms GPT-2 by one percentage point in terms of accuracy. It should be born in mind that these two models were tested on a binary case though, whereas the other ones were tested on a multiclass case. The models that performed the best on a multiclass case are C-Support Vector Classifier and BERT. To establish the absolute best classifier in a multiclass case, further research is needed that would deploy GPT-2 on a multiclass case.

1. Introduction

Text Classification is the procedure of designating predefined labels for text, and is an essential and significant part in many Natural Language Processing (NLP) tasks, such as sentiment analysis [1], topic labeling [2], question answering [3] and dialog act classification [4]. In the era that we live in, there are massive amounts of data and textual data is produced daily. Thus, it is highly inconvenient to process all this information manually. Moreover, due to fatigue or a lack of expertise, the accuracy of manual data processing is highly questionable. For these reasons, more and more people and institutions revert to automatic text classification to do the task with increased accuracy and reduced human bias. Distinction between shallow and deep learning models have been already investigated [4]. Mainly, shallow models dominated the text classification field since 1960s until the early 2010s. Shallow learning refers to statistics-based models, such as Naïve Bayes (NB), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). These methods had their fair share of success. However, they still need to do feature engineering, which costs time and financial resources. In addition, they disregard the natural sequential structure or contextual information in textual data. Thus, these models often fail to assign correct semantics to words. In this research paper, we test several different text classification models, some shallow and some deep.

2. Models

The shallow models tested in this paper are the well-explored Naive Bayes, Support Vector Machine and K-Nearest Neighbor. Bayesian classifiers assign the most likely class to a given example described by its feature vector [5]. On the other hand, the Support Vector Machine are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis [6]. Finally, the K-Nearest Neighbor is a non-parametric classification method, which is simple but effective in many cases. For a data record t to be classified, its k nearest neighbours are retrieved, and this forms a neighbourhood of t [7].

The deep neural models tested use Bidirectional Encoder Representations from Transformers (BERT) [8] and second generation Generative Pre-trained Transformer (GPT-2) [9], implemented by the Huggingface library [10].Both of them are transformers-architecture based models and differ fundamentally in that BERT has just the encoder blocks from the transformer, whilst GPT-2 has just the decoder blocks from the transformer. Moreover, GPT-2 is like a traditional language model that takes word vectors as input and estimates the probability of the next word as output. It is auto-regressive in nature: each token in the sentence has the context of the previous words. Thus, GPT-2 generates one token at a time[11]. By contrast, BERT is not auto-regressive. It uses the entire surrounding context all-at-once[12].

SwissText 2022: Swiss Text Analytics Conference, June 08–10, 2022, Lugano, Switzerland

ema.ilic9@gmail.com (E. Ilic);
m.garcia@pangeanic.com (M. G. Martinez);
m.souto@pangeanic.com (M. S. Pastor)

Table 1
Examples from the 'Text_Type' dataset

Text	Label
provisions relating to the act of accession of 16 april 2003	Legal
79 particulars to appear on the outer packaging	Medical
desloratadine was not teratogenic in animal studies.	Medical
there's no actress in town who can hold a candle to her.	Vernacular
each press of this button cycles through the following three indicator display options:	Tech
"a further leading interest rate indicator , the eurepo , was established in early 2002 ."	Finances

Table 2
Examples from the 'Digital' dataset

Text	Label
Averages over the reference period referred to in Article 2(2) of Regulation (EC) No 1249/96:	Email
A discount of 10 EUR/t (Article 4(3) of Regulation (EC) No 1249/96).	Marketing
"The risk is limited to the explosion of a single article."	Social Media
"a rating of 75 Ah, and "	Social Media

3. Methodology

The main idea of this research paper is to compare the results of different classifiers on two datasets, 'Text_types' and 'Digital', described later on in the section 3.1 with regards to relevant metrics, more precisely, precision, recall, accuracy, and F1.

3.1. Datasets

Two Pangeanic internal datasets are used for the experiments. The first dataset called 'Text_types' is comprised of 8.4M values and is divided into four classes: vernacular, legal, medical, tech and financial text. On the other hand, the second dataset is comprised of 1.3M values, and is representing digital text content divided into 3 classes: Social Media, Marketing and Email content. The second dataset is referred to as 'Digital'.

3.2. Tools

The experiments were executed using 24 parallelized CPU units of type x86_64, and NVIDIA Titan GPU with Cuda Version 11.0.

4. Experiments

Numerous different experiments, tests and trials have been done in order to observe the widest possible array of results. Namely, the codes have been executed with different quantities of data, on different servers, and on different datasets. While BERT was tested both as a multiclass as well as a binary model, GPT-2 was used as a binary model on all the classes of a certain dataset.

4.1. Case 1: Simple Classifiers and Grid Search

The 'Text_types' dataset was reduced to a total of 8437 units. The Randomized Search and Grid Search cross validation was applied with the help of scikit-learn library in order to choose the best hyperparameters for each simple classifier. The results are reported below. For the K-Nearest Neighbor, the optimal parameters chosen were the following: for the weights, the inverse weights with respect to the distance were choosen, and a total number of 3 nearest neighbors was chosen. On the other hand, the optimal parameters chosen in the grid search for Naive Bayes were a prior fit and the additive smoothing parameter was set to 0.01. The parameters chosen were Newton CG solver, no penalty and a constant was added to the decision function. For the C-Support Vector Classifier the kernel type chosen was 'rbf' and the degree of polinomial kernel function is 5.

4.2. Case 2: Binary BERT

The first model tested was binary BERT model by the 'huggingface' library. The 'text_types' dataset was reduced to 1687 samples for the sake of faster execution of the code. The dataset was turned into a binary one, in this case with 'legal' and 'non-legal' text categories. The analysis was conducted with pretrained BERT-base-uncased model and the results were the following: Namely, with this pretrained BERT-base-uncased model, the accuracy of 98.46% was achieved, accompanied by the f1 score of 98.72% for class legal and 98.07% for class non-legal.

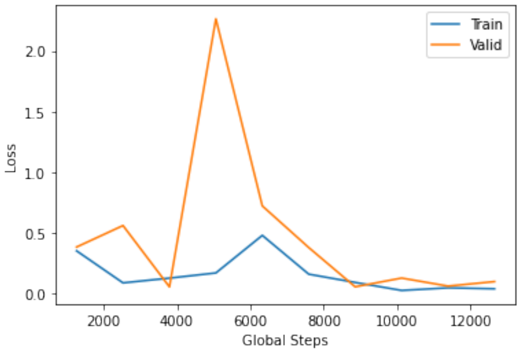

Figure 1: Training and Validation Loss for Legal vs. Non-Legal binary BERT

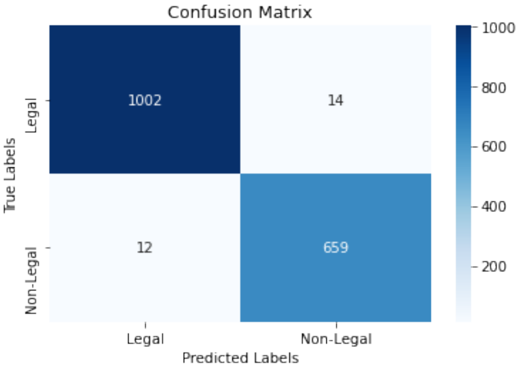

Figure 2: Confusion Matrix for Legal vs. Non-Legal binary BERT

4.3. Case 3: Multiclass BERT

On the other hand, BERT-base-uncased pretrained model was also used on a multiclass case of the same dataset ('Text_types') and later on the 'Digital'. The Text_types dataset was tested with 844 samples split into training and validation. The 'Vernacular' Class had the accuracy of 50/52, 'Finances' 10/14, 'Legal' 12/15, 'Medical' 24/26, and 'Tech' 18/20. The overall accuracy of the model on 'Text_types' dataset was therefore 89.76%.

For the 'Digital' dataset, on the other hand, 9000 samples were used which were later split to training and validation sets, and the BERT model was fine-tuned with the following results. The email, marketing and social media class had the true positive rates of 416/455 (91.43% accuracy), 417/456 (91.65% accuracy) and 425/455 (93.4% accuracy). Namely, this is a weighted accuracy of 92.05%.

4.4. Case 4: Binary GPT-2

Binary GPT-2 model by OpenAI was tested on 5062 samples of the 'Vernacular' vs 'Non-Vernacular' class of the Text_types dataset, with the weighted accuracy obtained of 98%. Below, one can observe the training and validation loss for the given classes as well as the confusion matrix.

The same model was also tested on all the three classes of 'Digital' dataset on a total of 13336 samples for training of each class.

As can be observed, the results for discriminating between the marketing and non-marketing class with the GPT-2 model were interesting, namely, a weighted average of 89% can be observed for GPT-2 trained on 'Digital' dataset. Below are visual representations of the success of this model on discriminating between the other two classes.

A weighted average of the accuracy between the social media and non-social media class was 96% and for the email vs. non-email class was 94%. The total weighted average accuracy of the binary GPT-2 model on the 'Digital' dataset was 93%.

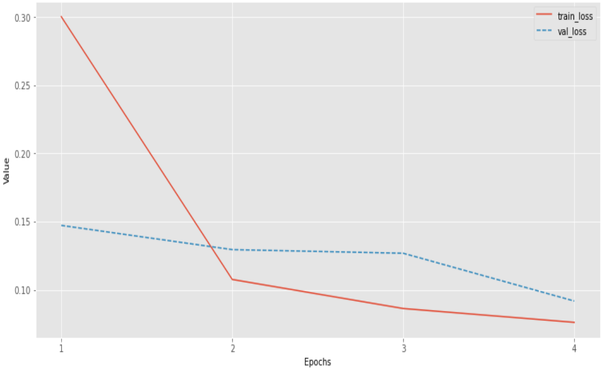

Figure 3: Training and Validation Loss for Vernacular vs. non-Vernacular class with GPT-2 on 'Text_type' dataset

4.5. Results

Results of the research may be observed in the Table 3. K-Nearest Neighbor, Multinomial Naive Bayes, Logistic Regression C-Support Vector Classifier and Linear Support Vector Machine Classifier were tested against the 'Text_Type' dataset, with the vectorization type chosen being Character level TF-IDF vector, whereas the Random Forest model was assigned the word level TF-IDF vectorization as the character one was incompatible with the classifier. The best results in terms of accuracy for the multiclass case were obtained

Table 3
Outcomes of Different Classification Models on 'Text_Type'

Classifier	Accuracy	Precision	Recall	F1
K-Nearest Neighbor	0.77	0.75-0.92	0.68-0.88	0.60-0.90
Multinomial Naive Bayes	0.89	0.81-0.93	0.72-0.96	0.77-0.94
Logistic Regression	0.89	0.76-0.94	0.79-0.93	0.79-0.93
C-Support Vector Classifier	0.90	0.83-0.99	0.74-0.99	0.78-0.95
Linear Support Vector Machine Classifier	0.88	0.78-0.92	0.80-0.96	0.80-0.95
Random Forest	0.78	0.55-0.92	0.65-0.92	0.60-0.88
BERT Pretrained Uncased	0.90	-	-	-
BERT binary (Legal/Non-Legal)	0.99	0.98-0.99	0.98-0.99	0.98-0.99
GPT-2 binary (Vernacular/ Non-Vernacular)	0.98	0.96-1.00	0.97-0.99	0.98

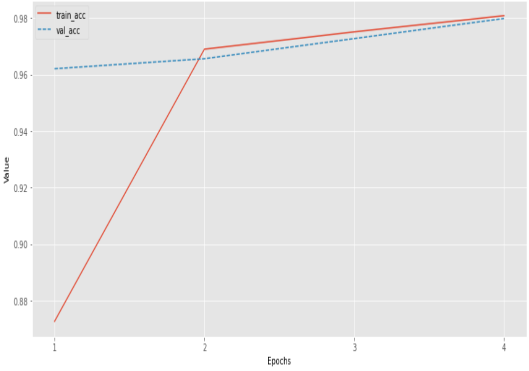

Figure 4: Training and Validation Accuracy for Vernacular vs. non-Vernacular class with GPT-2 on 'Text_type' dataset

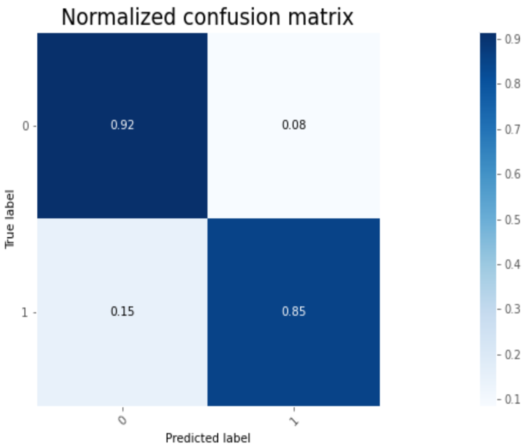

Figure 5: Confusion Matrix for Vernacular vs. non-Vernacular class with GPT-2 on 'Text_type' dataset

with the BERT model by the huggingface library and the C-Support vector classifier from the scikit-learn. On the other hand, the best results for the binary case were obtained with the GPT-2 classifier on a legal-vs non-legal class.

The absolute best results in terms of precision, recall and F1 were achieved for the binary BERT, whereas the best results in terms of those same metrics achieved for a multiclass case were by a C-Support Vector Classifier by the scikit-learn library. Bear in mind that the Precision, Recall and F1 for the BERT Pretrained Uncased remain unknown, and might indeed be greater than for the other classifiers.

5. Conclusions

According to our research, BERT and GPT-2 appear to perform excellent in a classification task, although BERT appears to be outperforming the GPT-2 by one percentage point in terms of accuracy. Both of these models significantly outperformed the shallow models, though it should be borne in mind that GPT-2 was only tested on a binary case. This is in line with the current research on the performance of large scale transformers models in classification tasks. [13] [14]

Some further research might be done comparing the performance of the multiclass GPT-2 on classification tasks in comparison to BERT. It would be interesting to observe if BERT always performs better, or if it only performs better on certain kinds of datasets.

References

[1] A. Maas, R. Daly, P. Pham, D. Huang, A. Ng, C. Potts, Learning word vectors for sentiment analysis, 2011, pp. 142–150.

[2] S. Wang, C. Manning, Baselines and bigrams: Simple, good sentiment and topic classification, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2:

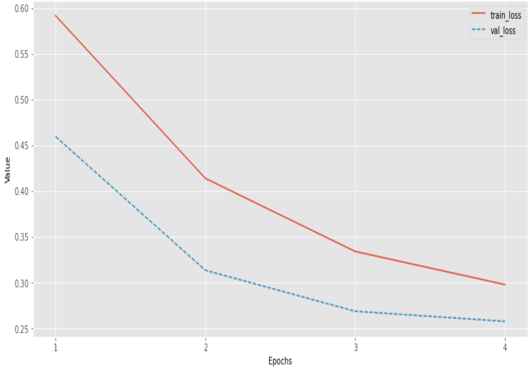

Figure 6: Training and Validation Loss for Marketing vs. non-Marketing class with GPT-2 on 'Digital' dataset

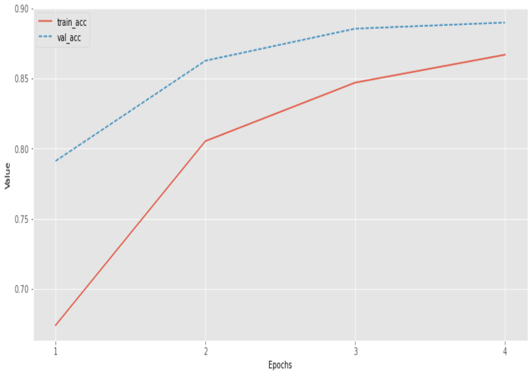

Figure 7: Training and Validation Accuracy for Marketing vs. non-Marketing class with GPT-2 on 'Digital' dataset

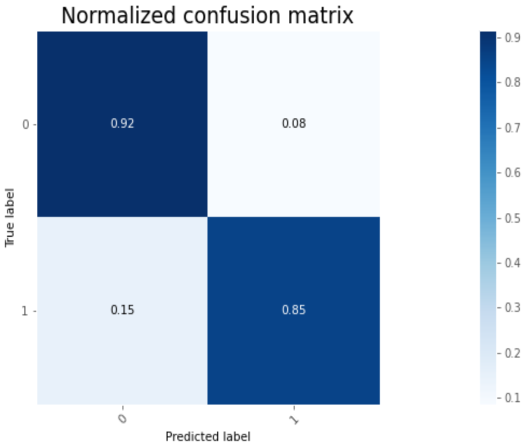

Figure 8: Confusion Matrix for Marketing vs. non-Marketing class with GPT-2 on 'Digital' dataset

Short Papers), Association for Computational Linguistics, Jeju Island, Korea, 2012, pp. 90–94. URL: https://aclanthology.org/P12-2018.

[3] Q. Mei, X. Shen, C. Zhai, Automatic labeling of multinomial topic models, in: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007, pp. 490–499.

[4] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, L. He, A survey on text classification: From shallow to deep learning, CoRR abs/2008.00364 (2020). URL: https://arxiv.org/abs/2008.00364. arXiv:2008.00364.

[5] I. Rish, An empirical study of the naïve bayes classifier, IJCAI 2001 Work Empir Methods Artif Intell 3 (2001).

[6] C. Cortes, V. Vapnik, Support vector networks, Machine Learning 20 (1995) 273–297.

[7] G. Guo, H. Wang, D. Bell, Y. Bi, Knn model-based approach in classification (2004).

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL: https://arxiv.org/abs/1810.04805. doi:10.48550/ARXIV.1810.04805.

[9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[10] Huggingface website, https://huggingface.co/, ???? Accessed: 2010-09-30.

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners (2018). URL: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[13] C. Sun, X. Qiu, Y. Xu, X. Huang, How to fine-tune BERT for text classification?, CoRR abs/1905.05583 (2019). URL: http://arxiv.org/abs/1905.05583. arXiv:1905.05583.

[14] S. González-Carvajal, E. C. Garrido-Merchán, Comparing BERT against traditional machine learning text classification, CoRR abs/2005.13012 (2020). URL: https://arxiv.org/abs/2005.13012. arXiv:2005.13012.