Harry PotBert or Bertmione Granger?
Machine Learning for Natural Language Processing 2020

Maxime Chabriel
Ensae Paris
maxime.chabriel@ensae.fr

Chloé Lavest
Ensae Paris
chloe.lavest@ensae.fr

April 24, 2022

Abstract

Is a BERT algorithm able to distinguish variations of writing styles from the expressed content ? Using JK Rowling's Harry Potter book series, we compare the performance of a BERT algorithm to a Bag of Word's in predicting the presence of a fiction character inside a random excerpt. We do not find significant evidence of the association of a writing *atmosphere* being associated with each of the characters. However, we believe further data treatment could lead to further results.

Problem Framing

The success of J. K. Rowling's worldwide bestseller Harry Potter book series has been the object of many a debate, arguments varying from the richness of the scenario to the complexity of the fictitious universe, or the writing style that is at the same time adapted to children, young and old adults. In this project, we especially want to study the mise en scene of the books' characters. Beyond a name and a personal lore, Rowling has without a doubt attributed to each of them a unique writing atmosphere : this ranges from speech mannerisms of the characters themselves, to the type of environment they are evolving in (dark, joyful, serene,. . .). To determine the importance of the use of such techniques by the author, we will resort to a predictive strategy : after having removed the names of a select few individuals inside Harry Potter excerpts, we will evaluate the performance of an algorithm to predict the missing information.

To do so, we build two different types of machine learning algorithm : one precise enough to understand a word in its context, and how it contributes to the meaning conveyed by the text, and another more holistic that groups words under similar lexical fields and contextual proximity. We then train these to predict the presence of main characters of the Harry Potter book series inside excerpts (paragraphs, as delimited by the author). Depending of the success of the algorithms to the task at hand, we thus have three possible outcomes. 1. The algorithms make similar predictions. If this is the case, there is no perceivable difference between a word sequence's meaning and the atmosphere it conveys. This has powerful implications as it would hint that distinction between the a text's *atmosphere* and the global meaning of the narration are not, in fact, two distinct objects. 2. The holistic algorithm and the contextual algorithm have similar performances, but on different types of excerpts. If this is the case, it shows that Rowling uses a two-steps writing, phases when she develops the atmosphere of the characters, working on image associations to build the characters in the head of the readers, and phases when she develops the narration of the series scenario. 3. One of the algorithms simply succeeds more than the other. In such a context, we could rule that an author is either defined by its place in the Harry Potter series scenario, or the image association Rowling built around each of them.

Data and methods

Our data is made of the 7 books of the Harry Potter saga[1], broken down into paragraphs (we use the paragraphs as they appear in the books, and we merge the paragraphs too small, such as dialogues, to have a database of observations with homogeneous size). We mask relevant character names using a grammar-based approach. We can see in Table 1 that the most recent books have more paragraphs that the others, and these also featuring more words on average (and median-wise). The standard deviation follows a much noisier pattern, which does not appear relevant[2]. As expected, the three most mentioned characters are Harry Potter, Ron Weasley and Hermione Granger (see Figure 1 in the appendix). This is true for all books, except the 6th one for Hermione, where Albus Dumbledore gets more citations. Looking at the most common words by book is not very informative, considering only very common English come up at the top, with only "Harry" ending up in the top 10 for every book.

[1] You can access the data processing notebook in our github at the following url: https://github.com/chlolv/NLP_Lavest_Chabriel.git

[2] Those results are also discernible in Figure 2

The unit of observation used for classifications is a paragraph of one of the books of the Harry Potter series. The label that will be predicted by the algorithm is the presence of one of the main characters of the series inside a paragraph. However, several characters can coexist in a similar paragraph, that contradicts the principle of a unique label for a unique paragraph. We thus resort to build compounded labels that match all the characters present in the excerpt.

For our holistic approach, we resort to use a Bag of Words vectorisation methodology (implemented with the Word2Vec[3] package). Proximity between words is established by their contextual proximity inside the corpus, and we represent each of these in a 100-dimensional space. We do not rely on a pre-trained Word2Vec model as a custom-trained algorithm allows to capture as much as possible variance inside our own corpus. Furthermore, as all words in a sentence do not have as much importance as the others, we weight the importance of these with their TF-IDF scores. Depending on the number of occurrences of a word in a targeted observation (a specific paragraph) and the whole corpus, the word will have a different importance in the building of the final vectorized representation of this observation. Building on this data, we then implement a Random Forest to predict the presence of the characters inside the paragraph. For this part of the project, the algorithm complexity was mainly driven by the Random Forest method, with a complexity of $O(ntree * dim * Nlog(N))$ (where ntree is the number of fitted trees, dim is the number of dimensions in which the words are projected unto by the Word2Vec, and N is the number of observations). The Word2Vec method, that follows an $O(N * log(V))$ complexity (N is the number of observations and V the size of the corpus vocabulary, excluding stopwords), is also a heavy method, given the size of our vocabulary.

For the task of contextual prediction, we resort to use a BERT[4] pre-trained model. BERT is a state of the art model that has proven to be highly performant in induction tasks. Thus, it makes a perfect candidate for a prediction relying on the meaning conveyed by the surrounding context to predict the presence of a character in an excerpt. This algorithm is very different from the holistic approach as it work with an *attention* mechanism that allows to isolate specific words inside a sentence. Thus, while it will be extremely efficient in manipulating the meaning of the words it is considering, it will fail to understand the whole meanings of the paragraph, making it impossible to capture the *atmosphere* that can radiate from it. Because of our limited computational capacity, we resort to only a two-linear-layers feed-forward classification fine-tuning of the pre-trained *base* bert model of the HuggingFace library. The complexity of our model is thus of $O(E * L_1 * L_2 * N)$ where E is the number of ran epochs, L_i is the size of the inner layer i and N is the number of observations.

Results

Both algorithms resulted in poor performances with an accuracy of 0.20 for the Random Forest (average F1-score of 0.19), and 0.17 for the BERT classification algorithm (average F1-score of 0.16)[5], which is rather close to a random prediction. Tests to increase scores resulted more than often in over-fitting. It is therefore hard to conclude anything in favor of one of our working hypothesis. One might argue that the predictive efficiency of the algorithms does not matter, as long as they are able to discern the pattern they should be discerning. However, comparisons of the predictions both of the BERT and the W2V × RF show that they do not identify any similar pattern in the data. This is clearly visible in the scatter plot (Figure 5) where we confront predictions from both algorithms.

Discussion/Conclusion

This work could be deepened by extending the number of characters taken into consideration, and widening their differences by focusing on individuals with more diverse statuses (such as Harry's adoptive members, other members of the Weasley family, character that only appear in few books...). The main limit of our analysis is the conundrum created by the situation where several characters appear at the same time. Is the resulting *atmosphere* a sum of each's ? Do they cohabit ? Or a new entirely different one emerges ? We opted for the third option, but this has the disadvantage of multiplying labels and thus diminishing the number of observations per category. A solution to such a problem could be to study each character one by one and run a different model on each one of them. Finally, the BERT algorithm could be used, instead of classification task, to do what he was originally supposed to do : predict masked characters in a context (we would be masking character names).

[3] [Mikolov et al.(2013)Mikolov, Chen, Corrado, and Dean]

[4] [Devlin et al.(2018)Devlin, Chang, Lee, and Toutanova]

[5] See both Figures (3 and 4) in the appendix

References

[Mikolov et al.(2013)Mikolov, Chen, Corrado, and Dean] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. Publisher: arXiv Version Number: 3.

[Devlin et al.(2018)Devlin, Chang, Lee, and Toutanova] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Publisher: arXiv Version Number: 2.

Table 1: Descriptive statistic of paragraphs by Harry Potter book

	HP 1	**HP 2**	**HP 3**	**HP 4**	**HP 5**	**HP 6**	**HP 7**
Number	1451	2140	2109	3600	5115	3134	3380
Avg. Length	54	40	50	54	51	55	59
Median Length	46	37	44	47	44	47	47
Std. Length	25	11	20	25	22	25	49

Table 2: Most frequent words in two selected Harry Potter books

HP 1		**HP 7**	
Word	*Count*	*Word*	*Count*
the	3309	the	9522
to	1844	and	5173
and	1807	to	4905
a	1580	of	4172
of	1248	a	3454
Harry	1206	he	2907
was	1172	Harry	2871
he	1033	was	2744
in	933	his	2493
his	895	it	2251

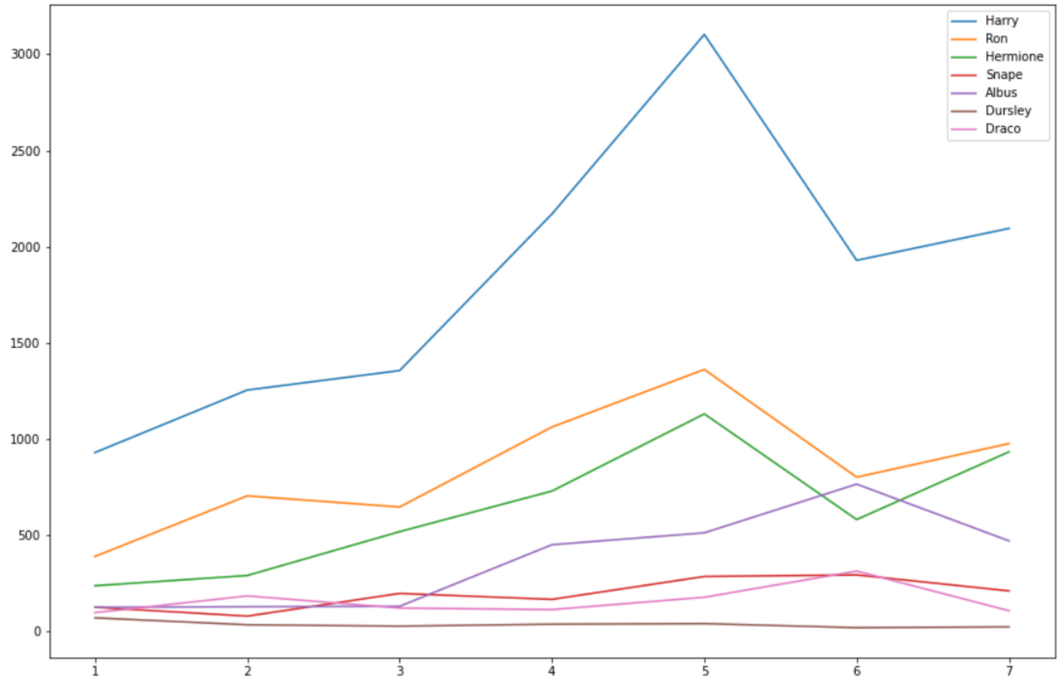


Figure 1: Citation of character by book

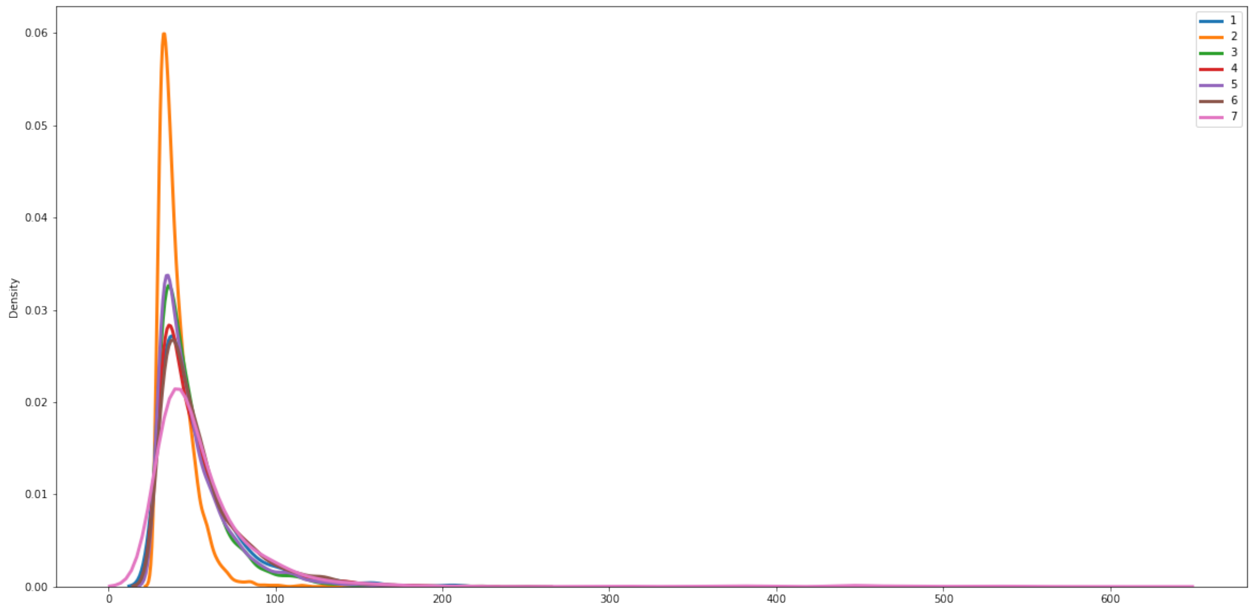


Figure 2: Distribution of paragraphs by book

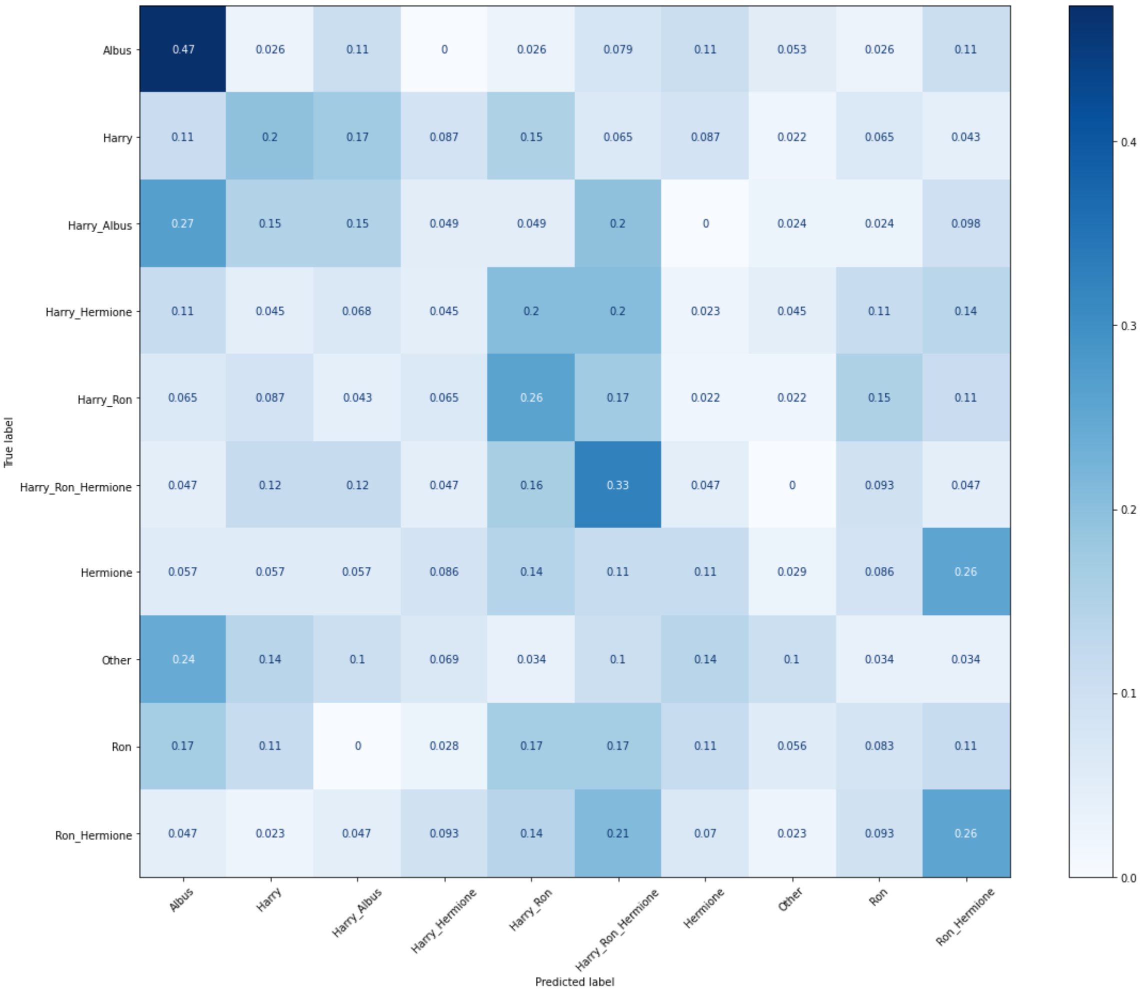
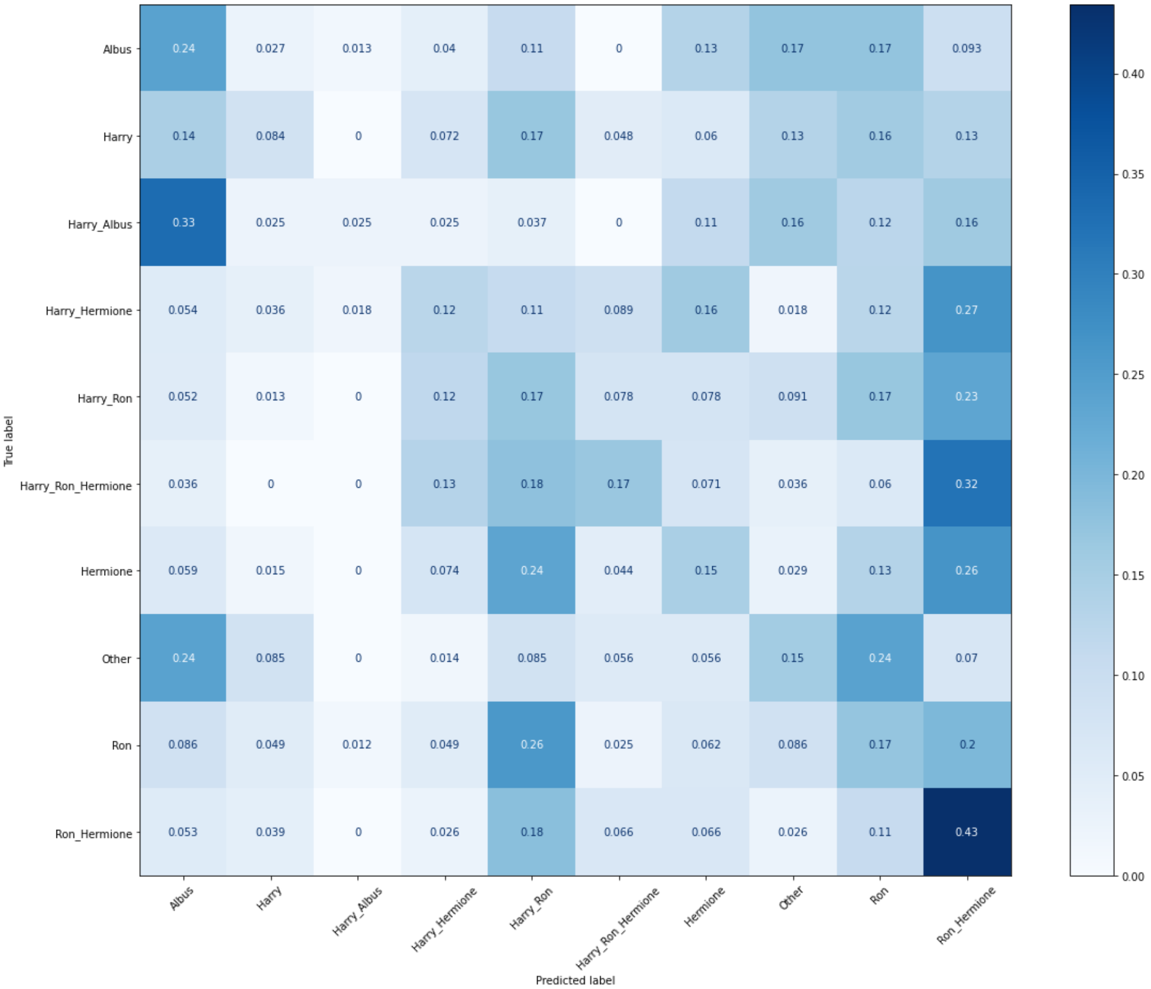


Figure 3: Confusion matrix of the Random Forest



	Albus	Harry	Harry_Albus	Harry_Hermione	Harry_Ron	Harry_Ron_Hermione	Hermione	Other	Ron	Ron_Hermione
Albus	0.24	0.027	0.013	0.04	0.11	0	0.13	0.17	0.17	0.093
Harry	0.14	0.084	0	0.072	0.17	0.048	0.06	0.13	0.16	0.13
Harry_Albus	0.33	0.025	0.025	0.025	0.037	0	0.11	0.16	0.12	0.16
Harry_Hermione	0.054	0.036	0.018	0.12	0.11	0.089	0.16	0.018	0.12	0.27
Harry_Ron	0.052	0.013	0	0.12	0.17	0.078	0.078	0.091	0.17	0.23
Harry_Ron_Hermione	0.036	0	0	0.13	0.18	0.17	0.071	0.036	0.06	0.32
Hermione	0.059	0.015	0	0.074	0.24	0.044	0.15	0.029	0.13	0.26
Other	0.24	0.085	0	0.014	0.085	0.056	0.056	0.15	0.24	0.07
Ron	0.086	0.049	0.012	0.049	0.26	0.025	0.062	0.086	0.17	0.2
Ron_Hermione	0.053	0.039	0	0.026	0.18	0.066	0.066	0.026	0.11	0.43

Figure 4: Confusion matrix of the Bert model

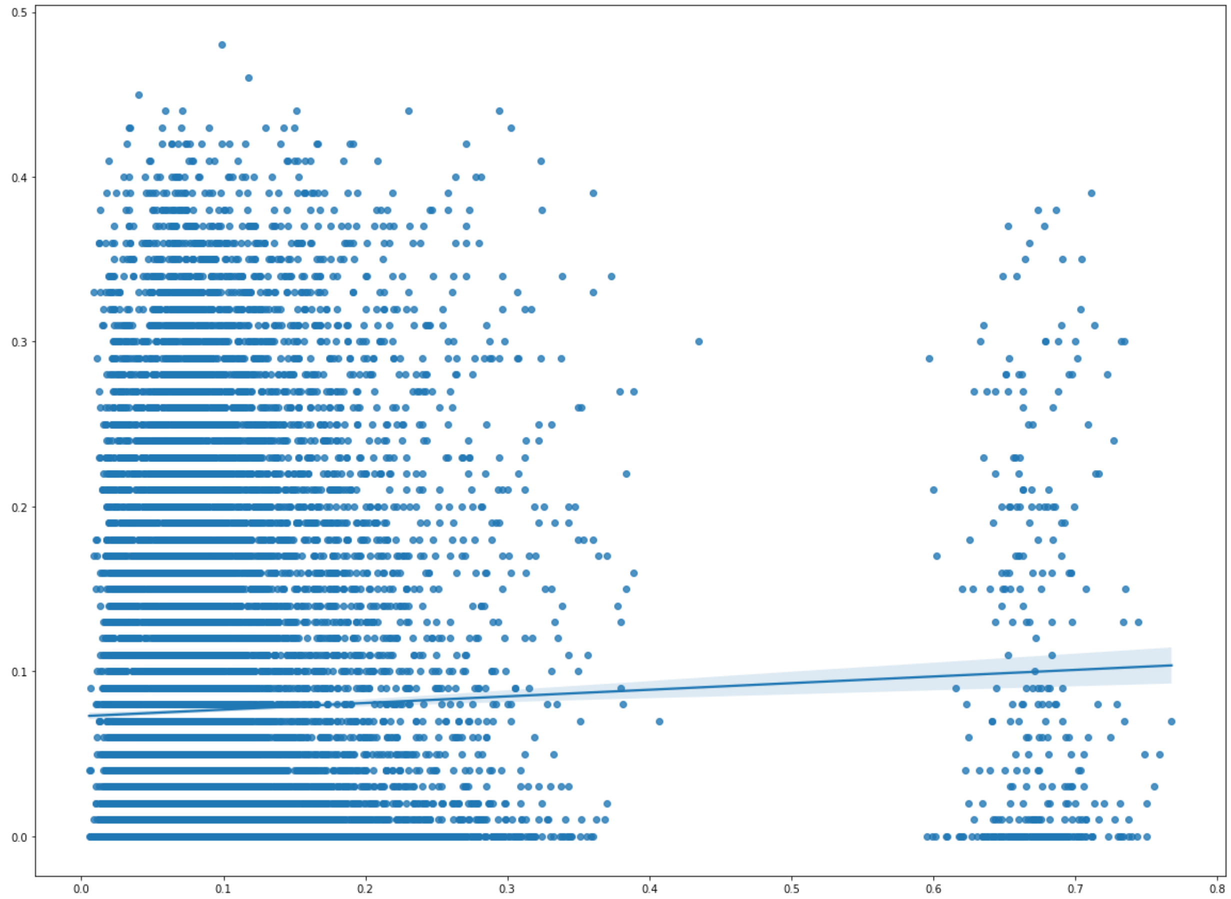


Figure 5: Bert (y axis) vs Random Forest (x axis) : Prediction of the label Harry Potter