**Explain in words the motivation behind this approach: imagine you don’t know the probability and statistics – how would you build the simplest classifier?**

I have to say that this is a difficult question. I think that, in one way or another, I’m biased from so many years of math and statistics. But if I have to think out of the box, I would classify a text looking at the words inside of it. If the text have “negative” words, it would be a negative review (for a sentiment classification). A few simple rules (like a decision tree), where the rules are constructed manually. Also I would use the length of the text and look if the words are complex words or not.

It’s really an interesting challenge.

**In the last task you implemented industry standard text preprocessing functions. Now, come back to the texts, read a bench of them and write a few words what a specific to this dataset transformation could be done in order to generalize the meaning of texts.**

First of all, I would delete these things -> “<br /><br />”

I would not remove stop words, they are useful for context. I would also not use any stemming or lemmatization function. But what I would do is to lower case the text and decontract possessive words like: “she’s”, “My dad’s” and all kind of contractions (‘re, ‘d, ‘ve, etc), I see a lot of them in the corpus.

I consider only these few transformations because I make the test and when the less transformations I made, the best the performance of the model. In case that I use my own trained embedding. I didn’t send the notebook because I have it in colab, where I use the free GPU in there to train it.

If is possible, I want to know the final answer to this question, it will be very helpful for me to continue learning. I know that I’m not a NLP specialist, but I would work hard to be one.