**Observations:**

* In this case, the models are slightly more accurate than a simpler rule like “always predict high”, since in our training sample 82% of reviews had the label high and our models had around 83% accuracy.
* However, our models do have better precision and recall than the simpler “always predict high” rule.
* The models with lemmatization and the model with lemmatization and stopword removal only show a small improvement in accuracy, while the one without text preprocessing is far better at recall.
* The biggest challenge was making the code efficient, since at first it was very slow. Once we had a working version, we used ChatGPTs help to make the code faster. The main difference is that we had to preprocess the text outside of the loop all at once, instead of processing each row one by one inside the loop.
* For our best predictions, we chose one of our iterations with the lemmatization and stopword removal model (the one highlighted on the table), since it was the most accurate (marginally) and it also had the best precision we did sacrifice a bit of recall, which probably means that we have either a smaller amount of true positives or a higher amount of false negatives. However, since precision is higher, we should have less false positives.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **‘k-fold’ iteration** | **Accuracy** | **Precision** | **Recall** | **F1** |
| no pre-processing | 1 | 0,837224 | 0,837871 | 0,997789 | 0,910864 |
| no pre-processing | 2 | 0,815111 | 0,815271 | 0,999245 | 0,897932 |
| no pre-processing | 3 | 0,828519 | 0,830037 | 0,99703 | 0,905902 |
| no pre-processing | 4 | 0,830363 | 0,830665 | 0,999259 | 0,907196 |
| no pre-processing | 5 | 0,814997 | 0,8147 | 0,999242 | 0,897584 |
| lemmatization | 1 | 0,840909 | 0,863576 | 0,960943 | 0,909662 |
| lemmatization | 2 | 0,826781 | 0,842867 | 0,967547 | 0,900914 |
| lemmatization | 3 | 0,840197 | 0,853611 | 0,974016 | 0,909847 |
| lemmatization | 4 | 0,830977 | 0,851537 | 0,964444 | 0,904481 |
| lemmatization | 5 | 0,827289 | 0,840656 | 0,971212 | 0,90123 |
| lemmatization and stopword removal | 1 | 0,842138 | 0,863276 | 0,963154 | 0,910484 |
| lemmatization and stopword removal | 2 | 0,820025 | 0,837696 | 0,966038 | 0,897301 |
| lemmatization and stopword removal | 3 | 0,837738 | 0,851852 | 0,973274 | 0,908524 |
| lemmatization and stopword removal | 4 | 0,829133 | 0,848958 | 0,965926 | 0,903673 |
| lemmatization and stopword removal | 5 | 0,82606 | 0,84 | 0,970455 | 0,900527 |