Is it necessary to read the entire document to classify?

Juan Martín Loyola[1,2*] and Marcelo Luis Errecalde[2]

1 IMASL, Universidad Nacional de San Luis, CONICET, San Luis, Argentina
2 LIDIC, Universidad Nacional de San Luis, San Luis, Argentina
* jmloyola@outlook.com

INTRODUCTION

The problem of classification in supervised learning is a widely studied one. Nonetheless, there are scenarios that received little attention despite its applicability. One of such scenarios is **early text classification**, which deals with the development of predictive models that can determine the class a document belongs to as soon as possible. Here a document is assumed to be processed sequentially, starting at the beginning and reading its containing parts one by one. In this context, it is desired to make predictions with as little information (as soon) as possible. The importance of this variant of the classification problem is evident in tasks like sexual predator detection, where one wants to identify an offender as early as possible.

It is important to note that the early text classification problem consists of two related and complementary tasks. On the one hand, the task of **classification with partial information**, which consists of obtaining an efficient predictive model when only partial information is available that has been read sequentially up to a certain point in time. The emphasis in this case is to determine which classification methods are more likely to achieve performance comparable to that obtained when classified using the entire document. On the other hand, we have the task of **decision of the moment of classification**, that is, in which point in time one can stop reading and classify with some degree of confidence that the prediction is going to be correct. [1]

OBJECTIVE

Here, we focus in the problem of classification with partial information, comparing the performance of different predictive models in the dataset R8 provided by Cachopo in [2]. We will like to know how much time (percentage of document read) does it takes to correctly classify most of the documents and find out which model does it earlier.

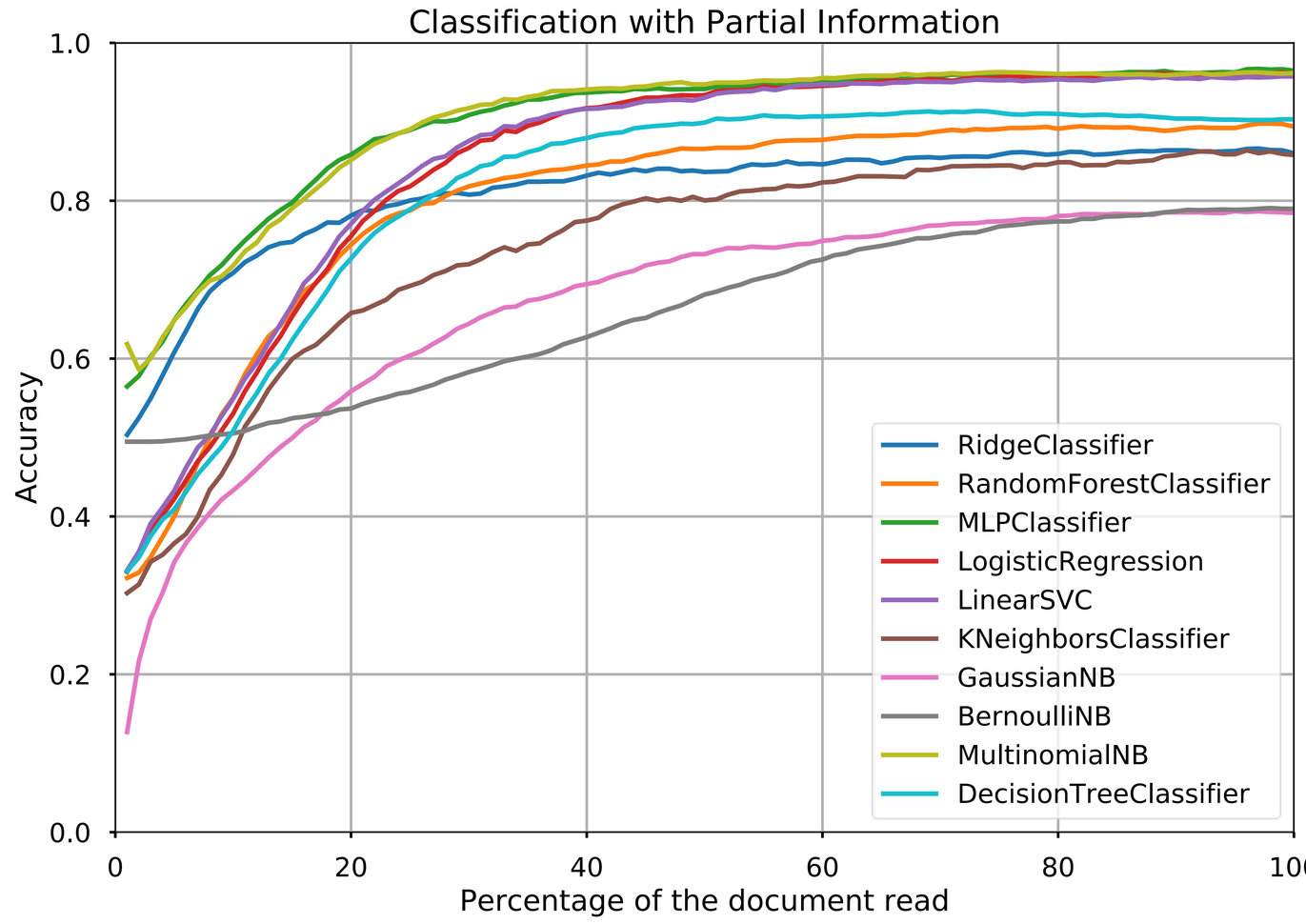
METHOD

In the task of classification with partial information we assume that during training we have full documents, therefore, the same training procedure as the standard supervised learning is performed. The difference comes at inference time: when classifying a new document we assume we read it in sequential order starting from the beginning (i.e. the first word from top to bottom and from left to right). This procedure was first devised in [3].

RESULTS

Distribution of documents in the R8 dataset

Class	# train docs	# test docs	Total # docs
acq	1596	696	2292
crude	253	121	374
earn	2840	1083	3923
grain	41	10	51
interest	190	81	271
money-fx	206	87	293
ship	108	36	144
trade	251	75	326
Total	**5485**	**2189**	**7674**



FUTURE WORK

- Implement different document representations, for example: word tf-idf, n-gram of characters tf, n-gram of characters tf-idf, n-gram of words tf, n-gram of words tf-idf.
- Evaluate this models in different kind of corpus, for example: the dataset of early risk prediction on the Internet (http://erisk.irlab.org), the Large Movie Review Dataset (http://ai.stanford.edu/~amaas/data/sentiment) and some of the other provided by Cachopo in [2].

REFERENCES

[1] Juan M. Loyola, Marcelo L. Errecalde, Hugo Jair Escalante, Manuel Montes y Gomez. Learning When to Classify for Early Text Classification. In Proc. of the 23rd Argentine Congress of Computer Science (CACIC 2017), pages 103-112, Universidad Nacional de La Plata, La Plata, Argentina.

[2] A. Cardoso-Cachopo. Improving Methods for Single-label Text Categorization. PhD thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa, 2007.

[3] H.J. Escalante, Montes-y-Gómez M., L.V. Pineda, and M.L. Errecalde. Early text classification: a naive solution. In Proc. of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@NAACLHLT 2016, pages 91–99, San Diego, California, USA, 2016.

FURTHER READING

Source code + digital version of poster (https://github.com/jmloyola/early-classification)