Chapter 15
Machine Learning Applications in Mega-Text Processing

Marina Sokolova
CHEO Research Institute, Canada

Stan Szpakowicz
University of Ottawa, Canada and Polish Academy of Sciences, Poland

ABSTRACT

This chapter presents applications of machine learning techniques to problems in natural language processing that require work with very large amounts of text. Such problems came into focus after the Internet and other computer-based environments acquired the status of the prime medium for text delivery and exchange. In all cases which the authors discuss, an algorithm has ensured a meaningful result, be it the knowledge of consumer opinions, the protection of personal information or the selection of news reports. The chapter covers elements of opinion mining, news monitoring and privacy protection, and, in parallel, discusses text representation, feature selection, and word category and text classification problems. The applications presented here combine scientific interest and significant economic potential.

INTRODUCTION

The chapter presents applications of Machine Learning (ML) to problems which involve processing of large amounts of texts. Problems best served by ML came into focus after the Internet and other computer-based environments acquired the status of the prime medium for text delivery and exchange. That is when the ability to work extremely large amounts of texts, which ML applications had not previously faced, became a major issue. The resulting set of techniques and practices, which we name *mega-text language processing*, are meant to deal with a mass of informally written, loosely edited text. A case in point is the analysis of opinions expressed in short informal texts written and put on the Web by the general public (Liu 2006). The sheer volume and variety of suddenly available language data has neces-

DOI: 10.4018/978-1-60566-766-9.ch015

sarily invited the use of computing software capable of handling such a mass of data, learning from it and acquiring new information.

Until now, no clearly delineated subfield of Natural Language Processing (NLP) dealt with mega-texts – textual data on the Web, computer-mediated text repositories and in general texts in electronic format. Text Data Mining – a form of Data Mining – concerns itself with deriving new information from texts, but most often restrains from the study of language. Still, many researchers focus on the study of language, for example lexical, grammar and style issues, in such texts (Crystal 2006; Liu 2006). That no overarching NLP discipline has emerged can be explained by the fact that electronic texts and old-fashioned texts in books or newspapers share major characteristics. We discuss these characteristics in the handbook chapter "Machine Learning in Natural Language Processing".

This chapter will show that ML techniques measure up well to the challenges that mega-texts pose. We focus on applications in aid of the study of language. In all cases which we discuss, an algorithm has ensured a meaningful result, be it the knowledge of consumer opinions, the protection of personal information or the selection of news reports. Although we mostly focus in this chapter on text classification problems, we go beyond document topic classification. English, the most popular language of the Web, is the default language of much of the scientific discourse. We state when problems deal with languages other than English.

In the chapter we cite standard measures used in NLP (Precision, Recall, F-score). Calculated for classifiers produced by an algorithm, they build on the numbers of correctly classified positive examples *TP*, incorrectly classified positive examples *FP*, and incorrectly classified negative examples *FN*.

$$\text{Precision: } P = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall: } R = \frac{TP}{TP + FN} \tag{2}$$

F-score is a weighted sum of Precision and Recall:

$$F = \frac{(\beta^2 + 1)P}{(\beta^2 + 1)P + R} \tag{3}$$

In some cases authors use the traditional Accuracy, which we cite:

$$\text{A} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

The remainder of the chapter is organized in four sections and some supplementary material. First, in three consecutive sections, we discuss ML applications in specific NLP mega-text problems. We concentrate on the correspondence between NLP problems and ML algorithms. The NLP problems arise from the need to analyze a high volume of electronic texts in marketing, mass media, and health care. Each problem is well served by different algorithms. Opinion analysis, originally a marketing problem, is mature enough to have become almost a traditional task in text analysis. On the other hand, privacy protection, especially urgent in health care, is the subject of cutting-edge research at the boundary of