Large text NLP Sex

Athens University of Economics and Business

Tsirmpas Dimitris, Gkionis Ioannis

November 20, 2022

1 Introduction

Understanding written text has always been an area of interest for computer research as well as commercial applications. Successfully parsing human text can be crucial for tasks traditionally necessitating lengthy human intervention, such as analyzing sentiments or extracting the underlying meaning and themes in works of literature. During the past few years, research has been pivoting away from more traditional statistical and machine learning methods in favor of emergent neural network and deep learning architectures to tackle this complicated problem. While these new methods are very promising, they face severe challenges when used in the context of very large documents such as books or scientific articles. In this paper we will examine many of these challenges, how they impact existing solutions and how recent methods in the field of NLP(Natural Language Processing) attempt to circumvent them. We will largely focus on three areas of scientific interest; Text categorization, which attempts to categorize documents into distinct classes, text summarization, which attempts to produce summaries of large documents and sentiment analysis, which attempts to identify the emotional tone behind a body of text.

2 Text Classification

Text classification refers to the process with which we can automatically categorize a document d in one or more classes c based purely on its contents. For example, a program used in a library would use text classification in order to assign labels such as ”Romance” and ”Science Fiction” to a new book.

Literary texts in particular are

Historically two methods were used for the task [1]. In the late 1980s knowledge engineering was used, according to which experts manually inserted a set of rules into the program, which were used to classify the documents. While this meant no training was necessary, the rigidity and human bias of manually inserting rules caused the programs to be unable to scale, generalize and adapt to new documents. Another problem that emerged is that a document could belong to multiple genres that could not be easily separated from each other.

A different approach uses machine learning classification algorithms such as Naive Bayes, SVMs (Support Vector Machines) and ID3 decision trees. Korde and Mahender [2] compile a list of most common machine learning approaches and briefly discuss their upsides, downsides and applications. Notably the K-NN (K-Nearest Neighbors) and Naive Bayes classifiers, while easy to compute and implement, are inefficient when features are highly correlated, Decision Trees can provide excellent insight for decision support tools at the cost of not being memory efficient for large bodies of text and SVMs are highly efficient at multiple categorization but require negative training sets which are harder to acquire. Furthermore they point to LLSF (Linear Least Squares Fit) as being ”one of the most effective text classifiers known to date”, albeit with a high computational cost.

References

[1] Rami Aly. “Hierarchical writing genre classification with neural networks”. bachelor. University of Hamburg, Oct. 18, 2018.

[2] Vandana Korde and C Namrata Mahender. “TEXT CLASSIFICATION AND CLASSIFIERS: A SURVEY”. In: *International Journal of Artificial Intelligence & Applications (IJAIA)* (Mar. 1, 2012).