

Large text NLP: Challenges and Solutions

Athens University of Economics and Business

Tsirmpas Dimitris, Gkionis Ioannis

November 26, 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.

2.1 Challenges

Text classification algorithms inevitably have to face challenges common to most tasks dealing with natural language processing. Some of these challenges are:

- **The curse of dimensionality.** The curse of dimensionality refers to the tendency for data to become exponentially sparse when projected to high-dimensional spaces compared to low-dimensional ones. Since grouping data is a fundamental procedure in statistics and machine learning, this means we are forced to use an exponentially large set of inputs for our methods to detect and organize the data in meaningful ways. More specifically, in the context of natural language processing, as our dictionary (the set of all unique words in our model) increases in size, the possible combinations of the words within it grow exponentially large, unlike our training samples.
- **Polysemy.** This refers to a word having multiple possible meanings. For example the word "bank" may mean a financial institution, a building or a synonym for relying upon someone. Our models must learn to differentiate the meaning of words depending on their context.
- **Homonyms.** Homonyms are words that either share the same spelling (homographs), the same pronunciation (homophones) or both. Using our previous

example the term "bank" as a financial institution and the term "bank" referring to a part of a river are homonyms (the difference between this and Polysemy is that homonyms are not related to each other semantically).

- **Sarcasm.** Sarcasm detection can be seen as a specific case of sentiment analysis, but is important to take account since it often completely inverts the meaning of a phrase.
- **Allegories.** Allegories are especially prevalent in literary works. Their challenge lies in their heavy use of metaphors, which flip the meaning our model has learned.

Literary texts in particular exhibit many additional challenges. First of all, their text is not structured and tends to communicate information in indirect and abstract ways, usually within complicated narratives. In practical terms this means that information crucial to determine the class of a book can be distributed across many chapters[6].

Another, more obvious problem, is the inputs' size. While CNNs (Convolutional Neural Networks) are considered a state-of-the-art technology, including in the field of Natural Language Processing, they grow proportionally to their inputs. This means that they are unable to process books that feed thousands of paragraphs as input without specialized and highly expensive hardware. Other neural network techniques specialized for NLP such as RNNs (Recurrent Neural Networks) and LSTM (Long-Short Term Memory) networks on the other hand struggle to conserve information over huge spans of text [5].

Furthermore, the language used in literary genres often drifts alongside the genre's own adaptations to social needs. This can lead to deficiencies if the focus of the classification program falls upon a small set of important words or sentences[4].

Finally, during our research we noticed that NLP models typically need to be trained on one language at a time, due to hardware/time constraints or because of their field of use. This might not be an issue for a model specializing in widespread languages such as English, French or Spanish, which contain extensive, supervised datasets of books and literary works, but not for less-spoken languages. This leads to research featuring such languages to favor less data-intensive models, such as decision tree classifiers ([4]), or artificially augmenting their datasets ([8]).

2.2 Early Work

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.

It's important to note that although new research has largely focused on the emerging applications of neural networks to solve the Text Classification problem, the traditional machine learning methods are still employed [4] [7]. In fact, tree classifiers and logistic regression can be competitive even in large literary texts [3].

2.3 Solutions

Most NLP implementations begin by restricting their model's dictionary by filtering non-statistically significant terms.[3] remove stop words, words that are too frequent/rare in respect to the entire corpus, words that are present in most classes, and choosing only the ones that are present in each book's individual dictionary when training. [6] on the other hand insist on keeping most words in the model's dictionary in order to enable their model to learn patterns that would otherwise be lost, such as tense and plurality.

Another practically necessary step is to limit how much of a document's content will be processed. As mentioned earlier, the computational complexity of neural networks, besides other challenges, prohibit training and testing models with the entire contents of books. [3] sample paragraphs for each document, while [6] use the 5000 first, last or random set of words for each document. Note that all these techniques (apart from random-5000) do not sacrifice the structure of the input's text.

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 (IJAI)* (Mar. 1, 2012).
- [3] Sicong Liu et al. "DeepGenre: Deep Neural Networks for Genre Classification in Literary Works". Language Technologies Institute, Carnegie Mellon University.
- [4] Monte Serrat, Mateus Tarcinalli Machado, and Evandro E S Ruiz, eds. *A machine learning approach to literary genre classification on Portuguese texts: circumventing NLP's standard varieties*. 2021.
- [5] Joseph Worsham. *Towards Literary Genre Identification: Applied Neural Networks for Large Text Classification*. 2014.

- [6] Joseph Worsham and Jugal Kalita. “Genre Identification and the Compositional Effect of Genre in Literature”. In: *Proceedings of the 27th International Conference on Computational Linguistics* (Aug. 20, 2018).
- [7] Baoxun Xu et al. “An Improved Random Forest Classifier for Text Categorization”. In: *J. Comput.* 7 (2012), pp. 2913–2920.
- [8] ΙΦΙΓΕΝΕΙΑ ΘΕΟΔΩΡΙΔΟΥ. “Ανάπτυξη και εφαρμογή μοντέλων γλώσσας σε ελληνικά λογοτεχνικά κείμενα”. ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ, 2020.