Aleksandrova Marina
NRU HSE

Comparative Analysis of Methods for Digital Representation of Textual Data in Social Sciences

It can be noted a high interest in the field of text mining over the years. This interest stimulates the development of methods of computer processing of text data, which opens new opportunities for various scientific fields.

Sociology researchers work closely with textual information, therefore introduction and development of new methods of working with text, and their adaptation and interpretation in particular considering the peculiarities of sociological science, is a significant research task.

In this paper, we consider one of the most important stages of textual data preprocessing - numerical representation of textual information. There are various methods for numerical representation of text data, and we will consider and compare the most common ones: “bag of words”[[1]](#footnote-1), TF-IDF (term frequency - inverse document frequency)[[2]](#footnote-2), BERT language model (Bidirectional Encoder Representations from Transformers)[[3]](#footnote-3). Each of the listed methods of digitizing text data embodies a certain approach, a set of rules and techniques, according to which numerical weights are assigned to elements of the text corpus (words, phrases, sentences, or whole texts), thus providing an opportunity for the statistical methods of analysis implementation on initially unstructured textual data.

Description and comparison of the listed methods of numerical representation of text data will be carried out according to the following plan. First, we will describe how each of the listed methods works. This will allow to understand how each of the methods works, and to analyze from the point of view of the goals of sociological research, what are the opportunities and limitations of these methods, what recommendations can be formulated for which research tasks, which methods of digitization will be more preferred. The same collection of texts can be represented in several ways, considering the needs of the researcher, and to solve different problems, he or she can refer to the results of text representation by different methods. It should be noted that this is not a common practice of working with text data, since, as a rule, the choice of method occurs mechanically - the quality metrics of machine learning models are compared, and the model that gives the highest indicators is chosen (the rest usually are swept aside, and no further study of the reasons for the quality differences is undertaken).

Secondly, we demonstrate, using the example of solving the practical problem of predicting the occurrence of item non-response depending on the wording of the questions in sociological questionnaires, how methods for numerical text representation work in terms of the quality of trained machine learning models, and compare these models.

As a summary, it will be shown what are the similarities and differences between the analyzed methods for numerical text representation, and recommendations will be formulated depending on the goals of researchers. The results obtained in this study contribute to the development of a methodology for analyzing textual information in social sciences, as well as a more conscious approach to working with such data.

1. Zhang Y., Jin R., Zhou Z.H. (2010) Understanding Bag-of-Words Model: A Statistical Framework. International Journal of Machine Learning and Cybernetics. No. 1. P. 43—52. [↑](#footnote-ref-1)
2. Hirschberg J., Manning C.D. (2015) Advances in Natural Language Processing. Science. Vol. 349. No. 6245. P. 261—266. [↑](#footnote-ref-2)
3. Devlin J. et al. (2018) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. URL: https://arxiv.org/pdf/1810.04805.pdf (Acessed: 22.10.2021). [↑](#footnote-ref-3)