UDC 000.04

VISUALIZATION OF SOCIAL TOXICITY TWEETS DISTRIBUTION

Aleksandr Kudriashov (ITMO University, faculty of informational communication technologies)

Academic advisor – docent of ICT faculty, doctor of technical sciences Maksim Khlopotov (ITMO University)

This report covers the following subjects: building a tool for automatic annotation of text messages from Twitter using Machine Learning; topic modeling; detection and visualization of topics distribution.

With the development of communicational technologies, it became possible to collect vast volumes of social and cultural data. Application of mathematical methods to collected data opens up opportunities for further sociological research of modern society, which is a trending direction of research in the whole world.

As a cultural phenomenon for analysis, the word “toxic” in social discourse was chosen. For example, “toxic relationship” means unhealthy relations, “toxic messages” – messages in which the author expresses a negative attitude towards the discussed subject. English-speaking segment of Twitter social network became the platform for data collection, with messages containing the word “toxic” being selected. The resulting dataset consists of 43 thousand tweets manually labeled into three categories: toxicity in medical terms, social toxicity, and unrelated tweets. Machine Learning algorithms are used for further automation of text annotation.

The following text preprocessing pipeline was established to prepare text data for the application of Machine Learning methods: noise reduction; tokenization; training, testing and validation dataset splitting. The training set is loaded to Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) architecture to create a model. The resulting model is being evaluated against testing and validation sets to ensure the quality of automatic annotation.

With the aid of the obtained model, it became possible to extract messages related to the category of interest from any time interval and to split data into topics. As a tool for topic modeling, Latent Dirichlet Allocation (LDA) was chosen, which uses mathematical statistics to distribute messages over a specified number of topics. The resulting number of messages distribution over topics is plotted against a timeline for patterns and anomaly detection.

As part of this work, set of text messages dated 2018 was prepared and neural network model was trained. The data was split into 15 topics using LDA, and for each topic the visualization was created. Results of visualization could be used for further sociological research.

Aleksandr Kudriashov (author) Signature

Maksim Khlopotov (academic advisor) Signature