ABSTRACT

Relevance of the research topic. Modern wide internet is a considerable source of data to be used in scientific and business applications. An ability to extract up to date data is frequently crutial for reaching necessary goals, though, modern quality solutions to this problem, which are using computer vision and other technologies, may be finantially demanding to acquire or develop, thus simple and cheap to develop, maintain and use solutions are necessary.

The purpose of the study is to create a software instrument aimed at extraction of structured data from news websites for usage in news trustworthiness classification. Following tasks were outlined and implemented to achieve the aforementioned goal:

- Outline existing approaches and analogues in areas of data extraction and news classification;
- Design and develop extraction, preparation and classification algorhitms;
- Compare the results achieved with developed extraction algorhitm and with existing software solution, including comparing machine learning accuracies on both of the extractors.

The object of the study is the process of text data extraction with subsequent machine learning analysis.

The subjects of the study are methods and tools of extraction and analysis of text data.

Scientific novelty of the obtained results. A simple greedy algorithm was created, combining the process of link discovery and data extraction. Expediency of usage of simple web data extraction algorithms for composing machine learning datasets was proven.

It was also proven that classical machine learning algorithms can achieve results similar to neural networks such as LSTM. Capabilities of machine learning systems to function efficiently in a bilingual context were also shown.

Publications. Materials, related to this study, were published in the All-Ukrainian Scientific and Practical Conference of Young Scientists and Students

“Information Systems and Management Technologies” (ISTU-2019) “News trustworthiness classification with machine learning”

WEB SCRAPING, WEB PAGE DATA EXTRACTION, CRAWLING, LINK DISCOVERY, MACHINE LEARNING, NEWS CLASSIFICATION, LONG SHORT-TERM MEMORY, NEURAL NETWORKS