STUDENTŲ ĮTRAUKIMO Į MOKSLINĘ VEIKLĄ SKATININAMOJO KONKURSO TEMA

Temos pavadinimas: Daugiaklasių tekstinių duomenų kokybės gerinimas naudojant didžiuosius kalbos modelius
Tikslas: pagerinti pasirinktų tekstinių duomenų kokybę klasių išsidėstymo prasme, siekiant pasiekti aukštesnį klasifikavimo algoritmų tikslumą.
Trumpas temos vykdymo aprašymas (ne daugiau kaip 2000 ženklų): Kuriant įvairios paskirties mašininio mokymo modelius reikalingi duomenys. Duomenys dažnai yra imami iš įvairių duomenų bazių, kurie kartais yra netinkamai paruošti, sužymėti ar pasitaiko kitų klaidų, todėl algoritmai negali išspręsti vienos ar kitos užduoties. Dažniausiai klaidos pasitaiko tekstiniuose duomenyse, kadangi tam tikras tekstas, sakinys ar pastraipa gali būti priskirta kelioms klasėms, o esama klasė būti netiksli. Atliekant šį tyrimą reikės: 1. Pasirinkti kelias tekstines duomenų aibes. 2. Naudojant didžiuosius kalbos modelius modifikuoti esamų duomenų aibių klases. 3. Pasirinkti skirtingus klasifikavimo algoritmus ir apmokyti juos naudojant pradines duomenų aibes ir modifikuotas. 4. Atlikti palyginamąją analizę išsiaiškinant didžiųjų kalbos modelio tinkamumą ir efektyvumą duomenų gerinimui. 5. Aprašyti tyrimo rezultatus.
Temą siūlantis mokslininkas/dėstytojas: doc. dr. Pavel Stefanovič

TOPIC OF A COMPETITION PROMOTING STUDENT ENGAGEMENT IN SCIENTIFIC ACTIVITIES

Topic: Improving the Quality of Multiclass Text Data Using Large Language Models
Goal: improve the quality of selected text data in terms of class distribution, in order to achieve higher accuracy of classification algorithms.
Short description (max. 2000 characters): Data is required to develop machine learning models for various purposes. Data is often taken from multiple databases, which are sometimes improperly prepared, labeled, or have other errors, so the algorithms cannot solve one or another task. Most often, errors occur in text data, since a specific text, sentence, or paragraph can be assigned to several classes, and the existing class may be inaccurate. In carrying out this study, you will need to: 1. Select several text datasets. 2. Modify the classes of existing datasets using large language models. 3. Select different classification algorithms and train them using the original and modified data sets. 4. Perform a comparative analysis to determine the suitability and effectiveness of the large language model for data improvement. 5. Describe the results of the study.
Supervisor researcher/lecturer: assoc. prof. dr. Pavel Stefanovič