1 priedas

STUDENTŲ ĮTRAUKIMO Į MOKSLINĘ VEIKLĄ SKATININAMOJO KONKURSO TEMA

Temos pavadinimas: Tekstinių duomenų kokybės gerinimas naudojant didžiuosius kalbos modelius
Tikslas: pagerinti tekstinių duomenų kokybę, įvertinant ar klasifikavimo uždavinyje naudojami duomenys pagerino modelio 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 chatGPT ar kitą didįjį kalbos modelį modifikuoti esamų duomenų aibių klases. 3. Pasirinkti skirtingus klasifikavimo algoritmus ir apmokyti juos naudojant pradines duomenų aibes ir modifikuotas. 4. Atlikti palyginamąją analizę išsiaiškinant didžiojo kalbos modelio tinkamumą ir efektyvumą duomenų gerinimui. 5. Aprašyti tyrimo rezultatus.
Temą siūlantis mokslininkas/dėstytojas: doc. dr. Pavel Stefanovič

THE TOPIC OF A COMPETITION PROMOTING STUDENT ENGAGEMENT IN SCIENTIFIC ACTIVITIES

Topic: Improving the quality of textual data using large language models

Goal: to improve the quality of the textual data by evaluating whether the data used in the classification task improved the accuracy of the model.

Short description:

Building multi-purpose machine learning models requires data. Data is often taken from various databases, which are sometimes improperly prepared, labeled or have other errors, so algorithms cannot solve one or another task. Most errors occur in textual data, as a certain text, sentence or paragraph can be assigned to several classes, and the current class may not be accurate. This study will require:

1. Select several text datasets.
2. Using chatGPT or another large language model to modify the classes of existing datasets.
3. Select different classification algorithms and train them using original data sets and modified ones.
4. Perform a comparative analysis to find out the suitability and effectiveness of the big language model for data improvement.
5. Describe the results of the study.

Supervisor: doc. dr. Pavel Stefanovič