Assignment No. 6

Title :-

perform proper data labelling operation on dataset.

Theory :-

Data labeling

Explore the uses and benefits of data labelling, including different approaches and best practices.

What is data labeling ?

Data labeling, or data annotation, is part of preprocessing stage when developing an ML model. It requires identification of raw data and then addition of one or more labels to that data to specify its context for models, allowing ML model to make accurate predications.

Data labeling underpins different ML and deep learning use cases, including computer vision & natural language processing (NLP).

How does data labeling work ?

companies integrate software, processes & data annotators to clean, structure and label data. This training data becomes foundation for ML models. These labels allow annotators to isolate variables with datasets and this, in turn enables selection of optimal data predictors for ML models.

Along with machine assistance, data labeling task require "human-in-loop (HITL)" participations. HITL leverages judgement of human "data labelers" toward creating training process by feeding models datasets are most applicable given project.

labeled data vs. unlabeled data

computer use labeled and unlabeled data to train ML models, but what is difference?

* labeled data is used in supervised learning whereas unlabeled data is used in unsupervised learning.
* labeled data is more difficult to acquire and store, whereas unlabeled data is easier to acquire and store.
* labeled data can be used to determine actionable insight whereas unlabeled data is more limited in its usefulness unsupervised learning methods can help discover new clusters of data allowing for new categorizations than labeling.

Data labeling approaches

Data labeling is a critical step in developing a high performance ML model. Though labeling appears simple, it's not always easy to implement. As result companies must consider multiple factor and methods to determine best approach to labeling. Since each data labeling method has its, scope duration of project is reduced.

* Internal labeling :-
using in house data science expert simplifies tracking provides greater accuracy and increase quality. However this approach typically requires more time & and favors large companies with extensive resources.

* Synthetic labelling :-
This approach generates new project data from pre-existing datasets which enhances data and quality and time efficiency. However, synthetic labelling requires extensive computing power, which can increase pricing.

* Programming labelling :-
This automated data labelling process uses scripts to reduce time consumption and rank for human annotation. However, possibility of technical problems requires HITL to remain as part of quality assurance process.

* More precise predictions :-
Accurate data labeling ensures better quality assurance within machine learning algorithm, allowing model to train and yield accepted o/p. otherwise, as old saying goes "garbage in, garbage out" properly labeled data provide ground truth for testing & iterating subsequent models.

* Better data usability :-
Data labeling can also improve usability

of data variables within a model. For example, you might reclassify a categorical variable as a binary variable to make it more consumable for model. Aggregating data in this way can optimize model by reducing number of model variables. Whether you're using data to build computer vision model or NLP models, utilizing high quality data is a top priority.

★ challenges :-

Data labelling is not without its challenges. In particular some of most common challenge are.

★ Expensive and time consuming :-

While data-labelling is critical for ML models, it can be costly from both a resource and time perspective. If business takes a more automated approach, engineering teams will still need to set up data pipelines prior to data processing & manual labelling will is almost always be expensive & time consuming.

★ Prone to Human error :-

These labeling approaches are also subject to human error which can decrease the quality of data. This in turn, leads to inaccurate data processing and modelling. quality assurance check are essential to maintaining data quality.

* outsourcing :-

This can be an optimal source for high level temporary project but developing and managing a freelance-oriented workflow can also be time-consuming. Through free lancing platforms provide comprehensive candidate information to ease vetting process, hiring unmanaged data labeling teams provides pre-vetted staff and pre-built data.

* crowd sourcing :-

This approach is quicker and more-effective due to its microtasking capability and web based distribution. However worker quality QA, and project management very serious crowdsourcing platform some of the most famous examples of crowdsourced data labelling is recaptcha. This project was two-fold in that it controlled for bots while simultaneously improving data annotation of images. For example, a recaptcha prompt would ask user to identify all photos containing car to prove that they were human and then this process could check itself based on results of other users. The input of from these users provided a database of labels of an array of images.

* Benifits and challenges of data labelling :-

The general trade off of data labelling is that while it can decrease a business time to scale, it tends to come at cost. Most accurate data

generally improves overall predictions. So despite its high cost value that it provides is usually well worth investment. Since data annotation provides more context to datasets, it enhances performance data analysis as well as ML and A.I applications. For example, data labeling produces more relevant search results across search engine platforms & other product recommendation can in e-commerce platforms. Lets delve deeper into other key benifits and challenges

Benifits :-

Data labeling provides users, team and companies with greater context, quality and usability. More specifically you can except.

* Data labeling best practices :-

No matter approach following best practicing optimize data labeling accuracy and efficiency.

1) Intuitive and streamlined task interfaces
2) consensus
3) label auditing
4) Transfer learning
5) Active learning

* Data labeling use cases :-

Through data labeling can enhance accuracy, quality and usability in multiple contexts across industries, its most prominent use cases include:

* computer vision
* natural language processing (NLP)

* IBM and Data labeling

IBM offers more resources to help transcend data labeling challenges and maximize your overall data labelling experience.

* IBM cloud annotations (Link resides outside IBM)
* IBM cloud object storage
* IBM watson

No matter your project size or timeline, IBM cloud and IBM watson can enhance your data training processes, expand your data classification efforts and simply complex forecasting model.

conclusion:

Thus, we have studied data labelling operations on dataset.