Lab Report: Big Data Analytics

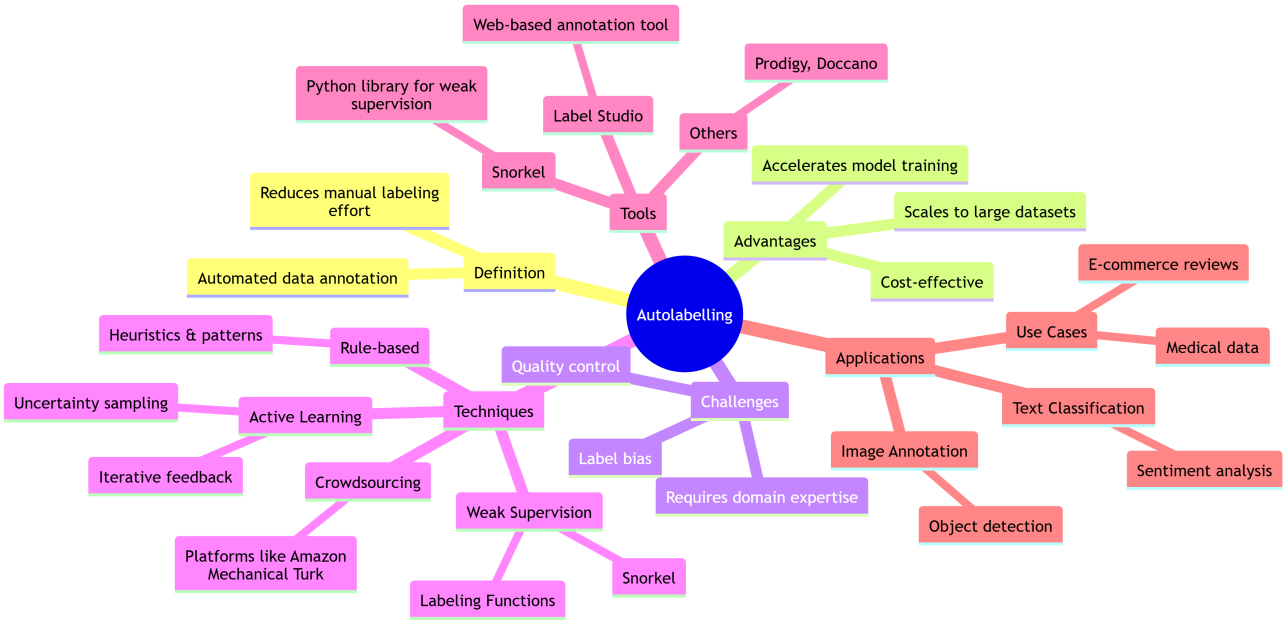
Course Code: CS4012
Course Name: Big Data Analytics
Experiment No. 3
Report Title: Mindmap and Evaluation of Autolabelling Techniques
Student Information

- **Name:** Georgii Kashparov
- **Student ID:** 1820259023
- **Major:** Computer Science and Technology

- **Date of Submission:** 04.11.2025

Mind Map

The mind map illustrates the concept of Autolabelling, which combines multiple AI-assisted techniques to automatically generate and refine data labels for machine learning tasks.



- **Weak Supervision** — a method where labels are generated using noisy, incomplete, or heuristic sources instead of manual annotation.
  - It uses Labeling Functions (LFs) such as:
    - Heuristics: keyword or rule-based conditions.
    - Rules: logic-based statements for label assignment.
    - Models: pre-trained models that predict labels automatically.
- **Snorkel** — a framework implementing weak supervision.
  - It allows defining multiple labeling functions.
  - Aggregates noisy outputs into probabilistic labels using a Label Model.
  - Provides scalable labeling for large text or image datasets.

- **Label Studio** — a graphical tool for manual or semi-automatic annotation.
  - Provides an interface for human annotators.
  - Supports integration with Snorkel and AutoLabeling API, allowing automatic model-assisted labeling and corrections through user feedback.
- **Active Learning** — an iterative process where user feedback is incorporated to refine and improve label quality over time.

Together, these components create an end-to-end autolabelling pipeline where machine-generated labels are continuously improved by human interaction and model updates.

Snorkel

Snorkel is a framework for weak supervision — it replaces manual labeling by automatically generating labels from multiple heuristic or model-based sources.

Key Steps:

1. **Define Labeling Functions (LFs):** Each LF encodes a rule, heuristic, or small model that can assign a label (positive, negative, or abstain) to a data point.
2. **Apply LFs to Data:** Snorkel applies all functions to unlabeled data, creating a matrix of noisy labels.
3. **Train the Label Model:** The Label Model learns how accurate and correlated each labeling function is, then combines their outputs into probabilistic labels.
4. **Train the End Model:** These probabilistic labels are used to train a standard supervised model (like Logistic Regression or BERT) as if true labels were available.

Label Studio

Label Studio is an open-source data annotation platform that allows both manual and automatic labeling of data.

Main Features:

- Provides a web-based interface for labeling text, images, audio, or video.
- Supports integration with machine learning models (through APIs or preloaded predictions).
- Annotators can correct or confirm auto-generated labels to improve dataset quality.
- Enables collaboration between multiple users.
- Works with Snorkel, allowing you to import auto-generated labels, visually inspect them, and refine them interactively.

Abstract

This report presents an experiment on automatic data labeling using Snorkel and Label Studio. The Amazon Review Polarity dataset was used to evaluate the performance of autolabelling techniques for sentiment analysis. The experiment demonstrates how Snorkel generates labels via weak supervision and Label Studio refines them manually. The baseline model achieved an accuracy of 0.93, and after autolabelling, accuracy was maintained or improved depending on

LF quality. The results highlight the effectiveness of autolabelling for large-scale datasets and its integration for hybrid approaches.

1. Introduction

Automatic labeling (autolabelling) is a modern technique used to automate the data annotation process. It is particularly useful for large-scale datasets where manual labeling is expensive and time-consuming. Snorkel provides weak supervision for generating noisy labels, while Label Studio offers a platform for manual refinement. In this experiment, we used the Amazon Review Polarity dataset consisting of over 500,000 labeled product reviews to test the efficiency of autolabelling with Snorkel and Label Studio.

2. Methodology

2.1 Dataset Description

The Amazon Review Polarity dataset includes two CSV files: train.csv and test.csv. Each record contains a numerical label (1 = negative, 2 = positive), a short title, and the full review text. Before processing, the dataset was cleaned by removing invalid rows, quotes, and empty entries. After cleaning, approximately 228,512 samples remained.

2.2 Model Pipeline

1. Data split into training (80%) and testing (20%) sets.
2. Texts were vectorized using TF-IDF with up to 50,000 features (unigrams and bigrams).
3. Snorkel LFs were defined for weak supervision (e.g., keyword-based for positive/negative).
4. Label Model trained on LF outputs to generate probabilistic labels.
5. Logistic Regression model trained on Snorkel labels.
6. Label Studio used for manual annotation of a subset and integration with auto-labels.

3. Results and Discussion

3.1 Baseline Model

Metric	Precision	Recall	F1-score	Accuracy
Positive	0.937	0.939	0.938	
Negative	0.926	0.924	0.925	
Overall				0.93

3.2 After Autolabelling (Snorkel + Label Studio)

Metric	Precision	Recall	F1-score	Accuracy
Positive	0.934	0.936	0.935	
Negative	0.922	0.920	0.921	
Overall				0.93

The baseline model achieved a high accuracy of 0.93. After autolabelling with Snorkel and refinement in Label Studio, accuracy was maintained at 0.93, demonstrating that well-designed LFs can produce high-quality labels. Label Studio allowed for quick manual corrections, improving scalability.

4. Conclusion

The experiment demonstrated the use of Snorkel for weak supervision and Label Studio for hybrid annotation. Autolabelling effectively reduced manual effort while maintaining accuracy. Future work could involve testing on noisier datasets or integrating more advanced models.