MINISTRY OF EDUCATION AND RESEARCH
TECHNICAL UNIVERSITY
OF CLUJ-NAPOCA, ROMANIA

FACULTY OF AUTOMATION AND COMPUTER SCIENCE
COMPUTER SCIENCE DEPARTMENT

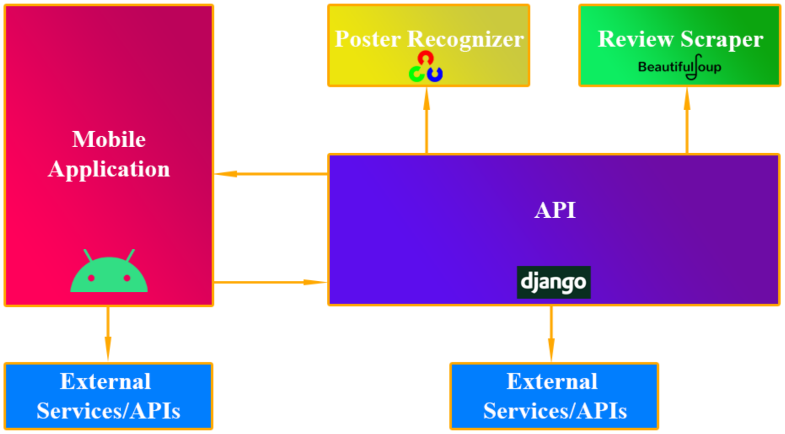
SUMMARY
of the License Thesis entitled:

REVIEW AGGREGATION APPLICATION USING WEB CRAWLING AND IMAGE PROCESSING

Author: **Paul Tudor LINCA**
Advisor: **Assist. Prof. Dr. Eng. Ion GIOSAN**

1. **Requirements:** Design and develop a system that provides the user a way to search for a desired movie, either manually or by taking a photo of its poster, and gathers its reviews from various sources. The application should also be able to provide movie recommendations based on the current day and the view history of the user.

2. **Proposed solutions:** The system is composed of numerous modules, each one providing certain functionalities:



- Poster recognizer: An OpenCV application, written in Python, that receives photos of movie posters and identifies and returns the title of the movie in question.
- Review Scrapers: A Python application that uses the Beautiful Soup library in order to scour the web for movie reviews given its title or identification code.
- External Services/APIs: Already existing systems with functionalities that benefit our project. Some of the functionalities provided by them are: authentication using external platforms, movie databases, poster databases.
- API: A web server implemented in Python using the Django framework that exposes the functionalities of the previous modules through its endpoints.

- Mobile Application: Android mobile application written in Kotlin that makes use of the API and other external services to achieve the overall purpose of the project. It also serves as the user interface of the system.

3. **Results obtained:** The resulted system consists of a native Android capable of displaying information about any movie. It makes use of web scraping techniques for review retrieval and image processing operations for movie image searching. External APIs are used for gathering movie information and a project specific API is used to enable communication between components through HTTP requests.

4. **Tests and verifications:** The application can retrieve information about 99% of the movies ever made. The web scraper can retrieve reviews for any movie as long as the sources have them available.
As the project's main functionalities are based on HTTP requests, the APIs used were tested and their response time measured. The performance of the external APIs was great, as expected. The project's API has a lesser performance due to the fact that optimization wasn't a priority. It was also deployed on the free version of Heroku, meaning that less resources are allocated to it. Nevertheless, it delivered satisfactory metrics.

5. **Personal contributions:**
   - Researched the possible technologies that can be used to achieve the project's purpose. I chose the latest and best documented tools available.
   - Learned two programming languages: Python and Kotlin.
   - Used new frameworks/libraries/tools: Android Jetpack / AndroidX, BeautifulSoup, Django and Heroku.
   - Developed an image processing application that detects and identifies movie posters from a photo.
   - Created two movie review web crawlers.
   - Created a web application that exposes the two previous applications through its endpoints**.** Deployed it on Heroku.
   - Created a state-of-the-art Android applications that uses all the best practices and latest tools and libraries.

6. **Documentation sources:** A part of the bibliographic research consists of:
   - Robert C. Martin, *Clean architecture: a craftsman's guide to software structure and design; Boston*, MA: Prentice Hall, 2018. ISBN: 9780134494166.
   - P. Shopa, N. Sumitha, P.S.K Patra, *Traffic sign detection and recognition using OpenCV*; published on the International Conference on Information Communication and Embedded Systems, 27-28, Feb. 2014. ISBN: 978-1-4799-3834-6

Date: 08.07.2020 Author: Paul Tudor LINCA

Coordinator: Assist. Prof. Dr. Eng. Ion GIOSAN