WEST UNIVERSITY OF TIMISOARA

DOCTORAL SCHOOL OF EXACT AND NATURAL SCIENCES

DOCTORAL THESIS ABSTRACT/REZUMAT

SUPERVISOR:
Prof. Univ. Dr. Marc FRÎNCU

PHD. CANDIDATE:
Elena FLONDOR

2025

WEST UNIVERSITY OF TIMIȘOARA

DOCTORAL SCHOOL OF EXACT AND NATURAL SCIENCES

COMPUTER SCIENCE

Machine Learning Methods for Identifying Mislabeled Applications and Android App Classification

Supervisor:
Prof. Univ. Dr. Marc Frîncu

PhD Candidate:
Elena Flondor

Doctoral Advisory Committee:

Prof. Univ. Dr. Dana Petcu, West University of Timișoara

Prof. Univ. Dr. Daniela Zaharie, West University of Timișoara

Conf. Univ. Dr. Ciprian Pungilă, West University of Timișoara

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Computer Science

2025

Abstract

Over the past decade, Android's rise as the leading mobile operating system has fueled massive application growth, with over 1.6 million on the Google Play Store. While its openness drives innovation, the sheer scale creates significant challenges in accurately classifying applications across categories. Accurate classification influences user experience and application discoverability, but mislabeled applications, which occur due to developer error or vague guidelines, can mislead users and skew analytics. Current systems rely heavily on developer input, making them unreliable and unscalable at today's application volumes. This thesis addresses two interrelated challenges: (1) detecting misclassified applications in existing datasets and (2) automating application classification using enriched feature sets and machine learning (ML) models. Using Latent Dirichlet Allocation (LDA) on a dataset of top applications, initial experiments confirmed that even highly ranked applications are sometimes inconsistently categorized. To address this, two novel methodologies for misclassification detection were introduced. The first approach employed semantic similarity measures combined with hierarchical clustering to identify applications that deviate from category norms. This method suggested the presence of mislabeled applications in the Medical and Weather categories. However, it was computationally expensive at scale. The second method improved efficiency and coverage by computing average K-Nearest Neighbor (KNN) distances using BERT-based representation of application descriptions. Potential mislabeled applications were validated through HDBSCAN clustering and summaries of cluster semantics generated by Large Language Models (Llama). Experiments across multiple datasets confirmed this technique's robustness, scalability, and interpretability. A significant contribution of the thesis is reversing the mislabeled detection process to identify highly representative applications for each category. This enabled the construction of well-aligned category benchmarks for developing and evaluating classification models. Building on this, the second significant contribution of the work focused on automating application and game classification using hybrid datasets enriched with static analysis (e.g., permissions, manifest components), dynamic features (e.g., GUI structure), and metadata from marketplaces (e.g., descriptions). Several ML models were evaluated, including traditional classifiers (Random Forest, Extreme Gradient Boosting) and a custom-designed multi-input Deep Neural Network (DNN). Experiments demonstrated substantial improvements over prior work. Random Forest achieved an average accuracy of 79% across multiple datasets, while Extreme Gradient Boosting improved classification quality when dynamic features were included. Using Llama-based embeddings of application descriptions significantly outperformed traditional text representations like TF-IDF, and the custom multi-input DNN model outperformed existing research: 94% accuracy, 88% precision, 90% recall, and an F1-score of 89%. These results highlight the strength of combining diverse feature types

with expressive data representations and ensemble learning strategies. Furthermore, the thesis validates the framework's generalizability by applying it to financial fraud detection, where adapted KNN-based outlier detection and ensemble models effectively identified fraudulent transactions, highlighting its flexibility and robustness beyond the Android domain. This thesis fills key research gaps and lays the groundwork for future advances, such as multi-label classification and behavioral analysis, ultimately supporting more reliable application marketplaces.

Rezumat

În ultimul deceniu, ascensiunea Android ca principal sistem de operare mobil a dus la o creștere masivă a numărului aplicațiilor, cu peste 1,6 milioane disponibile pe Google Play Store. Deși deschiderea platformei stimulează inovația, dimensiunea sa ridicată generează provocări semnificative în clasificarea corectă a aplicațiilor pe categorii. Clasificarea precisă influențează experiența utilizatorilor și descoperirea aplicațiilor, însă aplicațiile etichetate greșit, care apar din cauza erorilor dezvoltatorilor sau a unor ghiduri neclare, pot induce utilizatorii în eroare și pot denatura analizele. Sistemele actuale se bazează puternic pe inputul dezvoltatorilor, ceea ce le face nesigure și greu de scalat la volumele actuale de aplicații.

Această teză abordează două provocări interconectate: (1) detectarea aplicațiilor clasificate greșit în seturile de date existente și (2) automatizarea clasificării aplicațiilor folosind seturi de date îmbogățite și modele de învățare automată. Utilizând Latent Dirichlet Allocation (LDA) pe un set de date cu aplicațiile de top, experimentele inițiale au confirmat că și aplicațiile foarte bine cotate sunt uneori încadrate inconsistent. Pentru a remedia acest lucru, au fost introduse două metodologii noi de detectare a erorilor de clasificare. Prima abordare a utilizat măsuri de similaritate semantică combinate cu clustering ierarhic pentru a identifica aplicațiile care se abat de la normele categoriei. Această metodă a indicat prezența unor aplicații etichetate greșit în categoriile Medical și Weather. Totuși, s-a dovedit costisitoare din punct de vedere computațional. A doua metodă a îmbunătățit eficiența și acoperirea prin calcularea distanțelor medii K-Nearest Neighbor (KNN) folosind reprezentări bazate pe BERT ale descrierilor aplicațiilor. Aplicațiile potențial etichetate greșit au fost validate prin HDBSCAN clustering și generarea, folosind modele lingvistice de mari dimensiuni (Llama), a rezumatelor semantice ale clusterelor obținute. Experimentele pe mai multe seturi de date au confirmat robustețea, scalabilitatea și interpretabilitatea acestei tehnici. O contribuție semnificativă a tezei constă în inversarea procesului de detectare a erorilor de etichetare pentru a identifica aplicațiile foarte reprezentative pentru fiecare categorie. Acest lucru a permis determinarea unor seturi de aplicații bine aliniate pe categorii, utile pentru dezvoltarea și evaluarea modelelor de clasificare.

Pornind de aici, a doua contribuție majoră a lucrării s-a concentrat pe automatizarea clasificării aplicațiilor și jocurilor folosind seturi de date hibride, îmbogățite cu analiză statică (de exemplu, permisiuni, componente de manifest), caracteristici dinamice (de exemplu, structura GUI) și metadate din magazinele de aplicații (de exemplu, descrieri). Au fost evaluate mai multe modele de învățare automată, incluzând clasificatori tradiționali (Random Forest, Extreme Gradient Boosting) și o rețea neuronală cu multiple intrări, proiectată special. Experimentele au demonstrat îmbunătățiri semnificative față de lucrările anterioare. Random Forest a obținut o

acuratețe medie de 79% pe mai multe seturi de date, în timp ce Extreme Gradient Boosting a îmbunătățit calitatea clasificării atunci când au fost incluse caracteristicile dinamice. Utilizarea Llama pentru reprezentarea descrierilor aplicațiilor a depășit semnificativ reprezentările textuale tradiționale precum TF-IDF, iar rețeaua neuronală profundă cu multiple intrări a depășit rezultatele cercetărilor existente: 94% acuratețe, 88% precizie, 90% rată de regăsire și un scor F1 de 89%. Aceste rezultate evidențiază puterea combinării tipurilor de caracteristici diverse cu reprezentări expresive ale datelor și strategii de învățare de tip ansamblu.

În plus, teza validează generalizabilitatea cadrului propus prin aplicarea lui la detectarea fraudei financiare, unde adaptarea metodei KNN pentru detectarea anomaliilor și modelele de tip ansamblu au identificat eficient tranzacțiile frauduloase, demonstrând flexibilitatea și robustețea sa dincolo de domeniul Android. Această teză umple goluri cheie în cercetare și pune bazele unor cercetări viitoare, precum clasificarea cu etichete multiple și analiza comportamentală, ducând în final la magazine de aplicații mai fiabile.

Acknowledgments

I would like to sincerely thank my scientific advisor, Prof. Univ. Dr. Marc FRÎNCU, for his continuous guidance, support, and encouragement throughout my PhD journey and during the development of this dissertation. I'm genuinely grateful for his trust in me throughout this research, which allowed me to explore a research area I aim to pursue further in my career.

I also thank the members of the doctoral thesis guidance committee: Prof. Univ. Dr. Dana PETCU, Prof. Univ. Dr. Daniela ZAHARIE, and Conf. Univ. Dr. Ciprian PUNGILĂ, for their valuable insights and support.

A special acknowledgment goes to my family and boyfriend, whose unwavering support carried me through challenging moments and helped me remain committed to my goal. Their belief in my abilities and determination gave me the strength to complete this dissertation.

Finally, I offer my sincere gratitude to everyone who, directly or indirectly, contributed to the progress and completion of this scientific endeavor through their unconditional support.

Contents

List of Acronyms

AB AdaBoost.

AHC Agglomerative Hierarchical Clustering.

API Application Programming Interface.

ARFF Attribute-Relation File Format.

ASCII American Standard Code for Information Interchange.

Bayesian TAN Bayesian Tree-Augmented Naive Bayes.

BBN Bernoulli Naive Bayes.

BERT Bidirectional Encoder Representations from Transformers.

BIRCH Balanced Iterative Reducing and Clustering using Hierarchies.

BM25 Best Match 25.

BN Bayesian Network.

BoW Bag-of-Words.

CNN Convolutional Neural Network.

CPU Central Processing Unit.

CSV Comma Separated Values.

DBSCAN Density-Based Spatial Clustering of Applications with Noise.

DEX Dalvik Executable.

DL Deep Learning.

DNN Deep Neural Network.

DT Decision Tree.

ENN Ensemble Neural Network.

GNN Graph Neural Networks.

GPU Graphics Processing Unit.

GRU Gated Recurrent Unit.

GUI Graphic User Interface.

HDBSCAN Hierarchical Density-Based Spatial Clustering of Applications with Noise.

HTML Hypertext Markup Language.

JAR Java Archive.

KNN K-Nearest Neighbors.

LDA Latent Dirichlet Allocation.

Llama Large Language Model Meta AI.

LLM Large Language Model.

LR Logistic Regression.

LSTM Long-Short Term Memory.

MaxEnt Maximum Entropy.

MIME Multipurpose Internet Mail Extensions.

ML Machine Learning.

MNB Multinomial Naive Bayes.

MST Minimum Spanning Tree.

NB Naive Bayes.

NLP Natural Language Processing.

NMF Non-Negative Matrix Factorization.

NN Neural Network.

no number.

OOD Out-of-Distribution.

OS Operating System.

PLSI Probabilistic Latent Semantic Indexing.

RF Random Forest.

RNN Recurrent Neural Networks.

SDK Software Development Kit.

SGD Stochastic Gradient Descent.

SMOTE Synthetic Minority Oversampling Technique.

SMS Short Message Service.

SVM Support Vector Machines.

TF-IDF Term Frequency-Inverse Document Frequency.

UI User Interface.

UMAP Manifold Approximation and Projection.

URI Uniform Resource Identifier.

URL Uniform Resource Locator.

VSM Vector Space Model.

XGB Extreme Gradient Boosting.

List of Figures

List of Tables