



InnoCrowd, a product classification system for design decision in a crowdsourced product innovation

Indra Kusumah

► To cite this version:

Indra Kusumah. InnoCrowd, a product classification system for design decision in a crowdsourced product innovation. Computational Engineering, Finance, and Science [cs.CE]. Université Panthéon-Sorbonne - Paris I; Universität Konstanz, 2021. English. NNT : 2021PA01E050 . tel-03572748

HAL Id: tel-03572748

<https://theses.hal.science/tel-03572748>

Submitted on 14 Feb 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



UNIVERSITE PARIS 1–PANTHEON-SORBONNE

Indra KUSUMAH

STUDENT ID: 11544749

DEFENSE DATE: 10TH OF SEPTEMBER 2021

**InnoCrowd, a Product Classification System for Design
Decision in a Crowdsourced Product Innovation**

Prof. Dr. Anke HÖFFLER

Prof. Dr. Camille SALINESI

Prof. Dr. Josef WIELAND

Prof. Dr. Clotilde ROHLEDER

Prof. Dr. Hervé PANETTO

Prof. Dr. Dr.-Ing. Dr. h. c. Jivka OVTCHAROVA

Dr. Dianwicakasih ARIEFTIARA

Präsident

French Director

German Director

German Co-Director

French PhD-Reviewer

German PhD-Reviewer

Guest Jury

University of Constance

University Pantheon Sorbonne

Zeppelin University

University of App. Sci. Constance

University of Lorraine

University of Karlsruhe

University P.N. Veteran Jakarta

PREAMBLE

This PhD thesis describes my research as a PhD Student at the Centre de Recherche en Informatique (CRI) of University Paris 1-Panthéon-Sorbonne and at University of Applied Science Constance. It is under a cooperation program between those two institutions. Parallely with PhD work I am working as an IT Senior Solution Architect for more than 15 years in Germany, I helped customers implement the Siemens enterprise system Teamcenter for Product Lifecycle Management. Also I performed research to improve the industrial situations. Participation in these industrial projects and academic research conferences helped build an expertise network for promoting knowledge exchange. My network connections enabled contact between an industrial project leader at automotive industries such as BMW, Volkswagen, Jaguar Land Rover and my university research team.

My work as academic researcher at the Universities and as senior solution architect in the industry has enabled the publication of papers in international conferences in 24th ICE/IEEE International Technology Management in Stuttgart, in 9th International Conference on Computer Science and Information Technology (CoSIT 2020) in Zurich and in IFIP 18th International Conference on Product Lifecycle Management in Brazil.

THANKS

*I would like to express my deep thanks to Prof. Dr. **Camille Salinesi**, Professor at University Paris 1-Panthéon-Sorbonne and director of Centre de Recherche en Informatique for serving as my France PhD-director and giving me so much valuable advice during this research.*

*I would also express my highest thanks to Prof. Dr. **Josef Wieland**, Professor Emeritus of Zeppelin University, Germany for his endless help and patience while providing advice and reviewing papers and thesis, and for being my German PhD director which required travel from Germany to Paris to be present during my defense.*

*I would like to express my high thank to Prof Dr. **Clotilde Rohleder**, Professor at the University of Applied Sciences Constance, for serving as Co-Director and travelling from Constance, Germany to be present during my defense.*

*I thank particularly Prof. Dr. Dr.-Ing. Dr. h. c. **Jivka Ovtcharova**, Professor at University Karlsruhe, for participating in my defense as reviewer.*

*I thank Prof. Dr. **Hervé Panetto**, Professor at University Lorraine for serving on the jury as reviewer.*

I would like to thank colleagues and industrial partners who participated in many valuable discussions and supported my research in other ways. Their sharing issues and problems provided the basis for our work.

Last but not least, I would like to thank my wife, Zulfa Syafitri and my five childrens Afra, Hamzah, Salma, Zishan and Hilmi for their endless patience and support and pledge to be more available to them after this dissertation is finished.

ABSTRACT

System engineering focuses on how to design and manage complex systems. Meanwhile, in the era of Industry 4.0 and Internet of Things (IoT), systems are getting more complex. Contributors to higher complexity include the usage of modern components (e.g. mechatronics), new manufacturing technologies (e.g. 3D Print) and new engineering product development processes, e.g. open innovation. Open innovation is enabled by IoT, where people and devices are easily connected, and it supports development of more innovative products through ideas gained from predecessors and collaborators world wide. Some researchers suggest this approach is up to three times faster and five times cheaper than conventional approaches [Gassmann, 2012], [Howe, 2008], [Kusumah, 2018]. Because open innovation is relatively new, many managers do not know how to employ it effectively in some phases of product development [Schenk, 2009], [Afuah, 2017], including requirements definition, design and engineering processes (task assignment) through quality assurance. Also, they have trouble estimating and controlling development time and cost [Nevo, 2020], [Thanh, 2015]. As a consequence, the acceptance of this new approach in the industry is limited. Research activities addressing this new approach mainly address high-level and qualitative issues. Few effective methods are available to estimate project risk and to decide whether to initiate a project.

We propose InnoCrowd, a decision support system that uses an improved method to support these tasks and make decisions about crowdsourced engineering product development.

InnoCrowd uses natural language processing and machine learning to build a knowledgebase of crowdsourced product developments. InnoCrowd presents a manager with results of similar projects to show which practices led to good results. A manager of a new project can use this guidance to employ best practices for product requirements definition, project schedule, and other aspects, thereby reducing risk and increasing chances for success.

KEYWORDS: Innovation, Product Development, NLP, Machine Learning, Crowdsourcing, Requirements Engineering.

ABSTRAIT

L'ingénierie des systèmes se concentre sur la façon de concevoir et de gérer des systèmes complexes. En même temps, à l'ère de l'Industrie 4.0 et de l'Internet des objets (IoT), les systèmes deviennent de plus en plus complexes. Les facteurs contribuant à une plus grande complexité incluent l'utilisation de composants modernes (par exemple, la mécatronique), de nouvelles technologies de fabrication (par exemple, l'impression 3D) et de nouveaux processus de développement de produits d'ingénierie, par exemple l'innovation ouverte (Open Innovation). L'innovation ouverte est rendue possible par l'IoT, où les personnes et les appareils sont facilement connectés, et elle soutient le développement de produits plus innovants grâce aux idées acquises auprès de prédécesseurs et de collaborateurs du monde entier. Certains chercheurs suggèrent que cette approche est jusqu'à trois fois plus rapide et cinq fois moins chère que les approches conventionnelles [Gassmann, 2012], [Howe, 2008], [Kusumah, 2018]. Parce que l'innovation ouverte est relativement nouvelle, de nombreux gestionnaires ne savent pas comment l'utiliser efficacement dans certaines phases du développement de produits [Schenk, 2009], [Afuah, 2017], y compris la définition des exigences, la conception et les processus d'ingénierie (affectation des tâches) via l'assurance qualité. . De plus, ils ont du mal à estimer et à contrôler le temps et le coût du développement [Nevo, 2020], [Thanh, 2015]. En conséquence, l'acceptation de cette nouvelle approche dans l'industrie est limitée. Les activités de recherche portant sur cette nouvelle approche portent principalement sur des questions de haut niveau et d'ordre qualitative. Peu de méthodes efficaces sont disponibles pour estimer le risque d'un projet et pour décider s'il faut lancer un projet.

Nous proposons InnoCrowd, un système d'aide à la décision qui utilise une méthode améliorée pour prendre en charge ces tâches et prendre des décisions concernant le développement de produits d'ingénierie participative. InnoCrowd utilise le traitement du langage naturel (NLP : Natural Language Processing) et l'apprentissage automatique (Machine Learning) pour créer une base de données de connaissances sur les développements de produits issus ducrowdsourcing. InnoCrowd présente à un gestionnaire les résultats de projets similaires pour montrer quelles pratiques ont conduit à de bons résultats. Un responsable d'un nouveau projet peut utiliser ces conseils pour appliquer les meilleures pratiques pour la définition des exigences du produit, le

calendrier du projet et d'autres aspects, réduisant ainsi les risques et augmentant les chances de réussite du projet.

MOTS CLÉS: Innovation, Développement de produits, NLP, Machine Learning, Crowdsourcing, Ingénierie des exigences.

Résumé

Nos vies sont de plus en plus influencées par les technologies numériques. La numérisation change à la fois notre société et nous en tant qu'individus, y compris la façon dont nous travaillons et la façon dont les entreprises organisent leurs processus de travail et de création de services. Avec le crowdsourcing et le crowd work, un modèle innovant d'organisation du processus de création d'innovation de produits est proposé. On peut supposer que de plus en plus d'entreprises adoptent le concept afin de pouvoir accéder rapidement et de manière ciblée à un large pool de main-d'œuvre. À moyen et long terme, cela entraînera un changement disruptif dans les structures d'organisation et de travail.

Cependant, le changement n'est pas toujours aussi rapide que prévu, en particulier lorsque de nombreuses parties prenantes sont impliquées et que le processus d'exécution d'un développement de produit en crowdsourcing est également relativement complexe. Ce phénomène est prouvé par les résultats de recherche discutés dans ce travail. Une des tâches de ce travail est d'identifier les problèmes et défis liés à l'application du crowdsourcing dans le travail standard de l'organisation (entreprise de fabrication). Puis une autre tâche est d'identifier ce qui a été fait jusqu'à présent par les chercheurs pour résoudre ces problèmes.

Ce travail de thèse tente de contribuer à apporter une solution en développant une nouvelle méthode appelée InnoCrowd. Il s'agit d'un système d'aide à la décision basé sur l'intelligence artificielle pour le développement de produits en crowdsourcing. InnoCrowd devrait être en mesure d'aider les organisations à initier de nouveaux projets de crowdsourcing. L'organisation, qui n'est pas familière ou ne connaît pas les bonnes pratiques du Crowdsourcing, aura un accès relativement rapide aux informations sur les projets antérieurs liés, de sorte que le risque du projet soit mieux connu en tenant compte des expériences d'autres crowdsourcers dans le passé. InnoCrowd dispose d'un système de classification basé sur les connaissances, il est construit à l'aide d'algorithmes d'apprentissage automatique et de traitement du langage naturel. InnoCrowd travaille avec des banques de données provenant de données réelles de la plate-forme de crowdsourcing et entraîne l'algorithme d'apprentissage automatique avec ces données. L'ensemble des données est généré afin de construire le système de classification. Nous vérifions le concept de la solution en utilisant l'ensemble de données de la plateforme de crowdsourcing CAD Crowd (CAO Crowd) et TopCoder. Le résultat de la vérification de

l'analyse des exigences du produit (sous forme de fichier texte libre) donne un bon résultat de classification, à savoir un niveau de précision entre 90% et 91%. Par ailleurs, le résultat de la vérification de l'analyse du modèle numérique donne une précision encore plus élevée, entre 95% et 98%. La performance globale lorsque l'analyse du texte et de l'image est intégrée est de 90 %.

Afin d'optimiser d'avantage l'algorithme et d'éviter le sur-apprentissage, nous avons utilisé un algorithme génétique pour améliorer les paramètres de l'apprentissage automatique. Après l'amélioration avec l'algorithme génétique, InnoCrowd peut atteindre un niveau de précision de 95%. La comparaison d'InnoCrowd avec d'autres solutions donne un résultat clair : InnoCrowd est plus performant. En ce qui concerne l'utilisation d'InnoCrowd en tant que système d'aide à la décision, il est techniquement possible de fournir à l'utilisateur les informations relatives au projet précédent correspondant au projet prévu.

Trois cas d'utilisation industrielle utilisant des données réelles sont présentés. Le résultat de l'analyse d'InnoCrowd est montré. L'utilisateur aura accès aux informations sur le coût du projet précédent, le temps nécessaire pour terminer le projet et également les travailleurs du crowd qui ont travaillé sur une tâche similaire. Ces informations donnent une transparence claire sur le risque du projet, de sorte qu'InnoCrowd peut être utilisé comme réponse à la direction lorsqu'elle souhaite crowdsourcer un projet de développement de produit.

TABLE OF CONTENTS

CHAPTER 1	INTRODUCTION	13
1.1	PRODUCT DEVELOPMENT IN THE ERA OF GIG ECONOMY	13
1.2	SCOPE AND OBJECTIVES	15
1.3	RESEARCH METHOD	17
1.4	OVERVIEW OF THE CONTRIBUTION	18
1.5	DISSERTATION CONTENTS	19
CHAPTER 2	BACKGROUND AND DEFINITIONS.....	21
2.1	THE CREATION OF INNOVATIVE PRODUCTS	21
2.2	OVERVIEW ON CROWDSOURCING	23
2.2.1	<i>Definition of Crowdsourcing</i>	<i>23</i>
2.2.2	<i>Industrial application of crowdsourcing</i>	<i>23</i>
2.2.3	<i>The classification of crowdsourcing.....</i>	<i>25</i>
2.2.4	<i>Crowdsourcing process as an innovation platform.....</i>	<i>26</i>
2.3	MANAGEMENT OF CROWDSOURCING.....	27
2.3.1	<i>Crowd management in context of crowdsourcing</i>	<i>28</i>
2.3.2	<i>Task management</i>	<i>29</i>
2.3.3	<i>Contribution Management.....</i>	<i>30</i>
2.3.4	<i>Process / workflow management</i>	<i>31</i>
2.4	NATURAL LANGUAGE PROCESSING FOR TEXT ANALYSIS.....	32
2.4.1	<i>Word embeddings</i>	<i>32</i>
2.4.2	<i>Language modeling</i>	<i>33</i>
2.4.3	<i>Transformer models.....</i>	<i>34</i>
2.4.4	<i>Related application.....</i>	<i>34</i>
CHAPTER 3	PROBLEM STATEMENT	36
3.1	CROWDSOURCABILITY	36
3.2	PLATFORM.....	37
3.3	GAPS ALONG THE CROWDSOURCING PROCESS	40
3.4	RESEARCH HYPOTHESIS	41
CHAPTER 4	STATE OF THE ART ON PRODUCT CLASSIFICATION	42
4.1	PRODUCT CLASSIFICATION PROCEDURES	43
4.1.1	<i>Manual and automatic procedures.....</i>	<i>43</i>
4.1.2	<i>Statistical and non-distribution methods.....</i>	<i>43</i>
4.1.3	<i>Supervised and unsupervised procedures.....</i>	<i>44</i>
4.1.4	<i>Parametric and non-parametric methods.....</i>	<i>45</i>
4.2	SUPERVISED PROCEDURES FOR PRODUCT CLASSIFICATION.....	45
4.2.1	<i>Convolutional Neural Networks (CNN).....</i>	<i>45</i>

4.2.2	<i>Support Vector Machine (SVM)</i>	46
4.2.3	<i>Logistic Regression</i>	47
4.2.4	<i>Naive Bayes</i>	48
4.2.5	<i>Random Forest</i>	49
4.2.6	<i>Decision stump</i>	50
4.2.7	<i>AdaBoost</i>	50
4.2.8	<i>Related application</i>	51
CHAPTER 5	PROPOSED NEW APPROACH	53
5.1	GENERATING THE DATASET	53
5.1.1	<i>Collecting the dataset</i>	55
5.1.2	<i>Preprocessing of the dataset</i>	60
5.2	CONCEPT FOR TEXT ANALYSIS AND CLASSIFICATION	62
5.2.1	<i>Selecting the appropriate tool</i>	62
5.2.2	<i>Process chain for text analysis</i>	64
5.2.2.1	Corpus.....	65
5.2.2.2	Preprocess Text.....	66
5.2.2.3	Bag of Words	69
5.2.2.4	Machine learning algorithms for text analysis.....	70
5.3	CONCEPT FOR IMAGE ANALYSIS AND CLASSIFICATION	76
5.3.1	<i>Selecting the appropriate tool</i>	76
5.3.2	<i>Process chain for image analysis</i>	77
5.3.2.1	Import Images	78
5.3.2.2	Image Embedding	79
5.3.2.3	Machine Learning Algorithms for Image Analysis	80
5.4	INTEGRATION OF TEXT AND IMAGE ANALYSIS	81
CHAPTER 6	SCALE-UP THE CONCEPT OF INNOCROWD	83
6.1	METHODS FOR OPTIMIZING THE NEURAL NETWORK ARCHITECTURE.....	84
6.1.1	<i>Intuitive method</i>	84
6.1.2	<i>Construction methods</i>	85
6.1.2.1	Deconstructive methods.....	85
6.1.2.2	Constructive method	85
6.1.3	<i>Evolutionary methods (Genetic Algorithm)</i>	85
6.1.4	<i>Screenings</i>	86
6.2	OPTIMIZATION OF THE NEURAL NETWORK OF INNOCROWD WITH THE HELP OF A GENETIC ALGORITHM.....	86
6.2.1	<i>Basic configuration of genetic algorithm</i>	87
6.2.2	<i>Implementation of the genetic algorithm for InnoCrowd</i>	88
6.2.3	<i>The individual steps for the optimization of the network topology</i>	89
6.2.4	<i>The best NN parameter for InnoCrowd</i>	89
CHAPTER 7	INNOCROWD APPLICATION IN 3 REAL LIFE CASES	91
7.1	APPLICATION IN CONTEXT OF PRODUCT DESIGN	91
7.1.1	<i>Input data</i>	92
7.1.2	<i>Classification result done by InnoCrowd</i>	94
7.1.3	<i>Discussion on the result</i>	97

7.2	APPLICATION IN CONTEXT OF INFORMATION TECHNOLOGY TASK	99
7.2.1	<i>Input data</i>	99
7.2.2	<i>Classification result done by InnoCrowd</i>	101
7.2.3	<i>Discussion on the result</i>	103
7.3	APPLICATION IN CONTEXT OF ARCHITECTURAL DESIGN	104
7.3.1	<i>Input data</i>	105
7.3.2	<i>Classification result done by InnoCrowd</i>	106
7.3.3	<i>Discussion on the result</i>	109
CHAPTER 8 VALIDATION OF MAIN HYPOTHESIS AND REQUIREMENTS		111
8.1	REQUIREMENT SATISFACTION	111
8.2	HYPOTHESIS EVALUATION	112
8.3	COMPARISON OF EXISTING CONCEPTS.....	113
8.3.1	<i>Summary on Existing Concepts</i>	113
8.3.1.1	Concept from Thuan	113
8.3.1.2	Concept from Nevo.....	115
8.3.1.3	Concept of InnoCrowd	116
8.3.2	<i>Comparison Method</i>	117
8.3.2.1	ISO / IEC 9126	117
8.3.2.2	Boehm's Model	118
8.3.2.3	Similarity Measurement.....	119
8.3.3	<i>Applying the comparison method</i>	121
8.4	THREATS TO VALIDITY	126
8.5	OPEN ISSUES.....	126
CHAPTER 9 RESULTS AND CONCLUSIONS.....		127
9.1	SUMMARY	127
9.2	FUTURE WORK	129
REFERENCES		131

FIGURES

<i>Figure 1: Potential of Crowdsourcing in logistic sector.....</i>	<i>14</i>
<i>Figure 2: Scope of crowdsourcing (source: Clickworker.com)</i>	<i>16</i>
<i>Figure 3: InnoCrowd overview, system set up and system usage</i>	<i>18</i>
<i>Figure 4: Traditional vs crowdsourced task completion (source: Innosabi).....</i>	<i>21</i>
<i>Figure 5: From the age of manufacturing to the age of the customer (source: medium.com).....</i>	<i>22</i>
<i>Figure 6: Crowdsourced vehicle development in Local Motors</i>	<i>24</i>
<i>Figure 7: Crowdsourcing classification.....</i>	<i>25</i>
<i>Figure 8: Crowdsourcing process (source: Innosabi)</i>	<i>26</i>
<i>Figure 9: Main aspects to be managed in crowdsourcing</i>	<i>28</i>
<i>Figure 10: IT tools and platforms for open design</i>	<i>39</i>
<i>Figure 11: InnoCrowd overview, system set up and system usage</i>	<i>42</i>
<i>Figure 12: Supervised and unsupervised procedures.</i>	<i>44</i>
<i>Figure 13: Convolutional Neural Networks.....</i>	<i>46</i>
<i>Figure 14: Classification by using SVM.....</i>	<i>47</i>
<i>Figure 15: Logistic Regression</i>	<i>48</i>
<i>Figure 16: Naiv Bayes.....</i>	<i>49</i>
<i>Figure 17: Random Forest</i>	<i>49</i>
<i>Figure 18: Decision Stump.....</i>	<i>50</i>
<i>Figure 19: AdaBoost.....</i>	<i>51</i>
<i>Figure 20: Approach of InnoCrowd.....</i>	<i>53</i>
<i>Figure 21: Global CADCROWD designers and engineers.....</i>	<i>56</i>
<i>Figure 22: Dataset from a crowdsourcing platform (Source: CAD Crowd)</i>	<i>58</i>
<i>Figure 23: Images given by the crowdsourcer (Source: CAD Crowd).....</i>	<i>59</i>
<i>Figure 24: Contributions from the crowd (Source: CAD Crowd)</i>	<i>59</i>
<i>Figure 25: The collected text from the crowdsourcing platform.....</i>	<i>60</i>
<i>Figure 26: Collected image files</i>	<i>61</i>
<i>Figure 27: Orange – Data Mining</i>	<i>64</i>
<i>Figure 28: Process chain for text analysis</i>	<i>65</i>
<i>Figure 29: Each crowdsourcing project is represented as a single document in Corpus</i>	<i>66</i>
<i>Figure 30: Preprocessing text</i>	<i>67</i>
<i>Figure 31: Sample of the generated Bag of Words</i>	<i>70</i>
<i>Figure 32: Machine learning algorithms for text analysis.....</i>	<i>71</i>
<i>Figure 33: Layers of neural network.....</i>	<i>73</i>
<i>Figure 34: Comparison of machine learning algorithm performance.....</i>	<i>75</i>

<i>Figure 35: Process chain for image analysis</i>	78
<i>Figure 36: Import Images</i>	78
<i>Figure 37: Image Embedding</i>	79
<i>Figure 38: Machine Learning Algorithms for Image Analysis</i>	80
<i>Figure 39: Result of the image classification with deep learning Neural Network</i>	81
<i>Figure 40: Process chain for the integrated text and image analysis</i>	81
<i>Figure 41: Overfitting and underfitting</i>	83
<i>Figure 42: Genetic Algorithm</i>	86
<i>Figure 43: Optimization of the neural network architecture with the help of a genetic algorithm</i>	87
<i>Figure 44: Genotype and Phenotype</i>	88
<i>Figure 45: Sample of the product picture</i>	93
<i>Figure 46: The first related project to “Drone”</i>	94
<i>Figure 47: Images to the fist related project to “Drone”</i>	95
<i>Figure 48: The second related project to “Drone”</i>	95
<i>Figure 49: Images to the second related project to “Drone”</i>	96
<i>Figure 50: The third related project to “Drone”</i>	96
<i>Figure 51: Images to the third related project to “Drone”</i>	97
<i>Figure 52: The first related project to “Website Design”</i>	101
<i>Figure 53: The second related project to “Website Design”</i>	102
<i>Figure 54: The third related project to “Website Design”</i>	103
<i>Figure 55: Possible design of Floor Plan</i>	106
<i>Figure 56: The first related project to “Floor Plan”</i>	106
<i>Figure 57: Images to the fist related project to “Floor Plan”</i>	107
<i>Figure 58: The second related project to “Floor Plan”</i>	107
<i>Figure 59: Images to the second related project to “Floor Plan”</i>	108
<i>Figure 60: The third related project to “Floor Plan”</i>	108
<i>Figure 61: Images to the third related project to “Floor Plan”</i>	109
<i>Figure 62: Conceptual model developed by Thuan</i>	114
<i>Figure 63: Design to crowdsource</i>	114
<i>Figure 64: Process Design</i>	115
<i>Figure 65: A Road Map for Crowdsourcing in the Organizational Context</i>	116
<i>Figure 66: InnoCrowd’s conceptual model</i>	117
<i>Figure 67: ISO / IEC 9126</i>	117
<i>Figure 68: Boehm’s model</i>	119
<i>Figure 69: Quality Profile of Final Product for ISIC Entertainment and Report Tool Product</i>	121

TABLES

<i>Table 1: Crowdsourcing support from various software development platforms.....</i>	<i>38</i>
<i>Table 2: ISO / IEC 9126 in detail.....</i>	<i>122</i>
<i>Table 3: Comparison between InnoCrowd and Thuan's solution.....</i>	<i>125</i>

Chapter 1 INTRODUCTION

1.1 Product Development in the Era of Gig Economy

The digitalization era gives enormous impact to the society and enables the emerging of new trend: application of digital business model in complex product manufacturing, e.g., automobile industry. Hereby the collaboration format between the stake holders is different from the common way. The main difference is that the creativity task (product development) is outsourced to internet community [Achi, 2015]. By doing this they are able to reduce for example a car development time and cost enormously.

The growing number of people connected through internet is steadily increasing, which impacts both personal and industry practices. In this industry context, this trend recently has - among others, enabled crowdsourcing, that consists in “obtaining information or input into a task or project by enlisting the services of a large number of people, either paid or unpaid, typically via the internet” (source: Oxford dictionary). Jobs are offered on the web. People can work at home and submit results online. The distinction of crowdsourcing to outsourcing is that the tasks are offered open to the crowd and not to a certain group. Crowdsourcing allows anyone to participate, from less qualified to professional person. Crowdworkers are considered as independent contractors rather than employees, they are not guaranteed minimum wage. Crowdsourcing uses collective intelligence to solve the problem. The relation between crowdsourcing and collective intelligence is analog to a process which uses the appropriate tool for handling the task.

Crowdsourcing will be a game changer for the industry. In logistic sector it will make the sector more competitive, and major trucking companies could be at a risk of losing \$310 billion of operating profits to players enrolled on crowdsourced platforms. By helping smaller firms raise utilization levels, they will make the industry more competitive, and bring societal benefits such as logistical cost reductions for customers of \$800 billion. They could also reduce CO2 emissions by 3.6 billion tons, while generating additional income for consumers who decide to use their personal vehicles to deliver goods while on personal trips [Weforum, 2021].

Figure 1 shows the superiority of crowdsourcing when compared with other services in logistic sector.

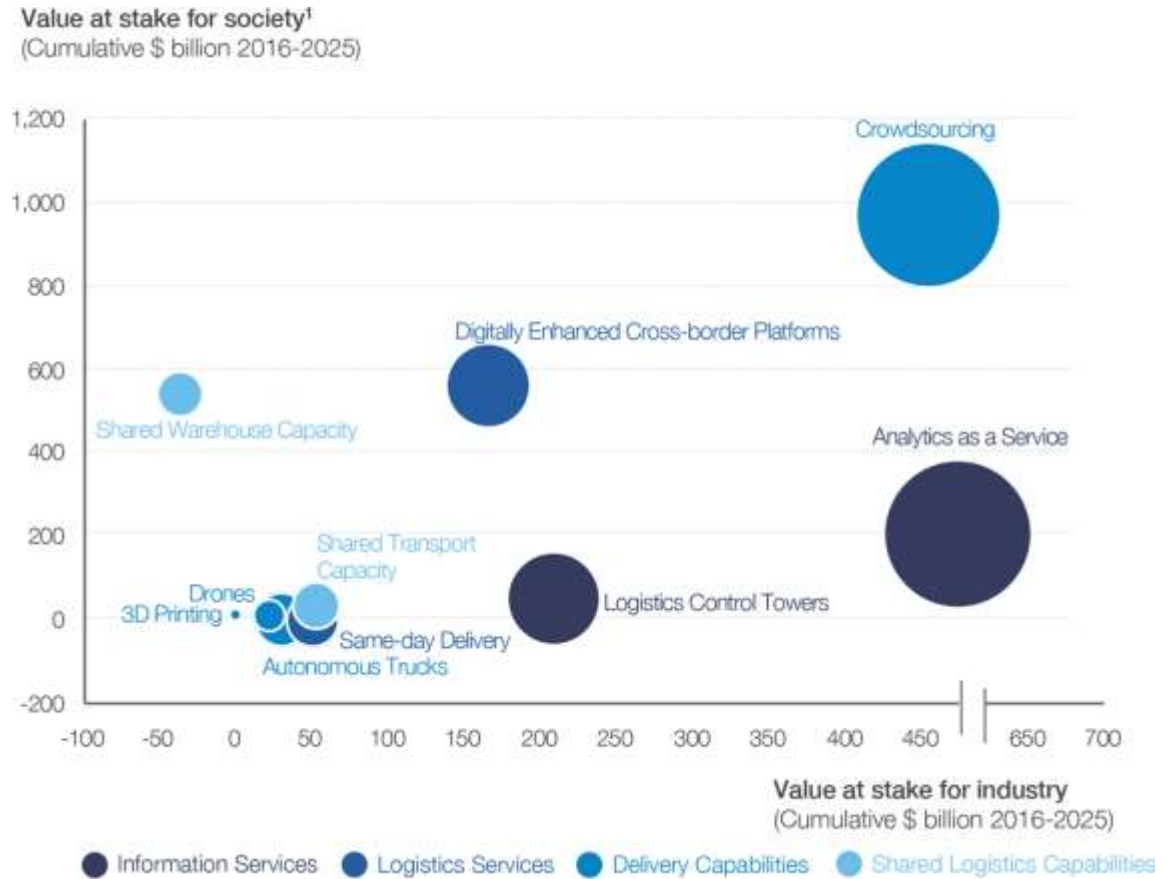


Figure 1: Potential of Crowdsourcing in logistic sector

Crowdsourcing applies to several kinds of activities, such as marketing, R&D, engineering, etc. Jobs are offered on the web. Freelancers execute them at home. “Crowdsourcing” and “Gig – Economy” are the terms related to this new trend. The gig economy encompasses all sorts of arrangements for non-employees performing work, but generally three types of stakeholders are involved [Li, 2020]:

- Requester: a company who defines and gives the task and later as a user of the result
- Workers (crowd): independent worker from internet as task executer
- Crowdsourcing Platform: Act as a connector between the user and the crowd.

A leading crowdsourcing company is Amazon Mechanical Turk, which occupies 2.676 workers (crowd) for performing 3,8 million tasks on their platform. The Raspberry PI project did not pay the crowd for much of the development of single-board minicomputer.

Compared with traditional approaches, crowdsourcing provides the following advantages:

- Lower cost and faster production [Eagle, 2009]
- Flexibility to select different workers for each project based on their project-related competences [Kittur, 2011].
- Strong involvement of the potential customer in the engineering process that increases the value of the end product [Gassmann, 2012]

1.2 Scope and Objectives

Figure 2 gives an overview about Crowdsourcing, which have six possible scopes. “Crowdtesting” and “Crowdfunding” are obviously far away from our work scope. We will create a product and not new knowledge or creative design for example for fashion or product logo, therefore “Collaborative Knowledge” and “Creative” are also not our scope. The task given in our crowdsourcing project is not a microtask such as tagging a picture or translating a text. Our work scope is the “Open Innovation”, where the crowds provide innovative product ideas.

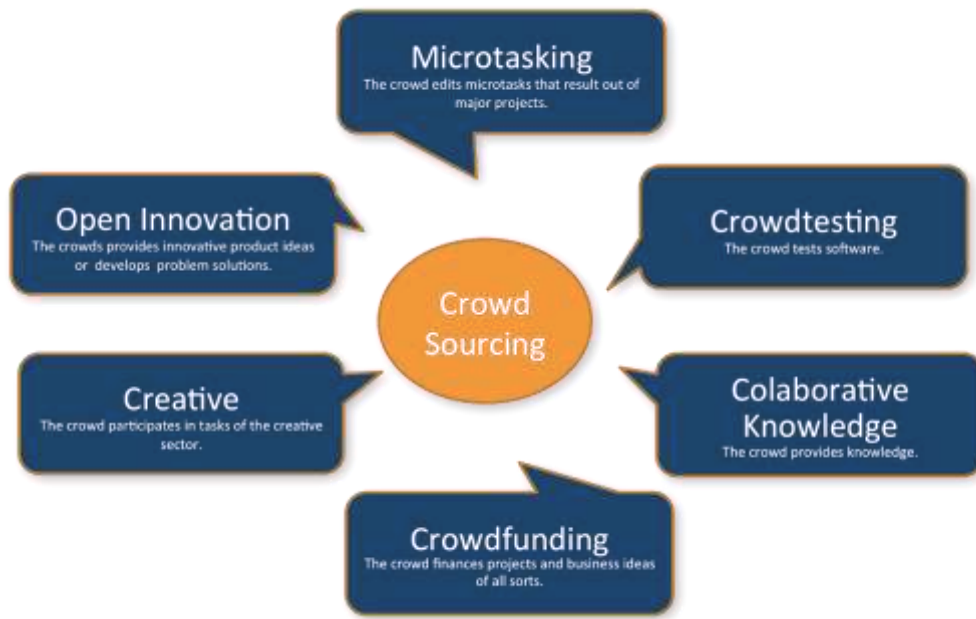


Figure 2: Scope of crowdsourcing (source: Clickworker.com)

We also limit the kind of product to be developed, namely only engineering product and software. From the type of data to be analyzed we limit the scope into only processing text and image data.

This PhD work can be seen as a continuation of research done by Achi [Achi, 2016], Rohleder [Rohleder, 2009] and Salinesi [Salinesi, 2016]. Achi [Achi, 2016] addressed the effect of the digitalization to the industries on how the way they create new innovation. It is basically the same starting point of the research project. And then this PhD work analyzed product requirement in detail for the classification of the possible output product. This is the same approach in general and comparable with the work from Rohleder [Rohleder, 2009], where she classified product requirement with methods further developed by her. And then in InnoCrowd there is an engine built up, which is a knowledge-based system. The similar approach was performed by Salinesi [Salinesi, 2016] where he built up a reusable knowledge-based system in requirement engineering.

The objectives of this PhD work are:

- To explain why some companies are hesitant to adopt crowdsourcing
- To determine typical crowdsourcing problems

- To identify shortcomings of the existing solutions for crowdsourcing problems
- To develop a new solution which has a better performance than the existing solutions

1.3 Research Method

We use “Design Science” as our research methodology [Peffer, 2012]. We collect system requirements, define acceptance criteria, and consider relevant research projects in order to produce research contribution rather than just routine design. We take real data from the crowdsourcing platform CAD Crowd.

Typically crowdsourcing tasks and product requirements are given in text and images.. Therefore, the first requirement in our research methodology is how to assist a manager in fetching all related text and image data effectively and easily. This task can be fulfilled by a web scraper.

The second requirement is how to prepare the data for automated analysis, including deletion of data that is irrelevant for the classification task. Here we can use text and image editors.

The third requirement is to develop a scalable algorithm for the classification of text and image data. We can consider Natural Language Processing (NLP) for text data. Through an experiment we describe in a later chapter, we will select or develop a machine learning algorithm to handle image processing.

The next step is the definition of the acceptance criteria. The most important one is the level of accuracy for the data classification result, which should be 90%. Others include time, cost, and quality.

According to Vallantin [Vallantin, 2018] the accuracy between 60% and 70% is poor, accuracy between 70% and 80% is good and accuracy between 80% and 90% is excellent. This work is aiming at accuracy of 90% so that the result is highly qualified.

We conducted a comprehensive literature study to determine gaps in the state of the art. It is shown by mentioning how it applies to each aspect of this work.

1.4 Overview of the Contribution

We developed a novel method for the generation of a knowledgebase relevant to a specific crowdsourcing provider. Figure 3 shows an overview of InnoCrowd including system set up and system usage. The block "System set up" summarizes how we built the Decision Support System (DSS) [Kusumah, 2018]. Decision support implies the use of computers to:

- Assist managers in their decision processes in semi structured tasks.
- Support, rather than replace, managerial judgement.
- Improve the effectiveness of decision making [Tripathi, 2011]

Development of a DSS starts with the collection of data from the crowdsourcing provider, which consists of text files (product requirements from the product customer), images, project histories, and digital models (2D and 3D files of the product to be designed). The second step uses NLP and a Deep Learning Neural Network (DLNN) to extract information. The third step embeds product classification in the Neural Network (after NN training and test). The last step establishes relationships between each product sub class and the project specific information relevant for a decision: product specification, related time and cost, and a description of the working team from the crowdsourcing provider's project history.

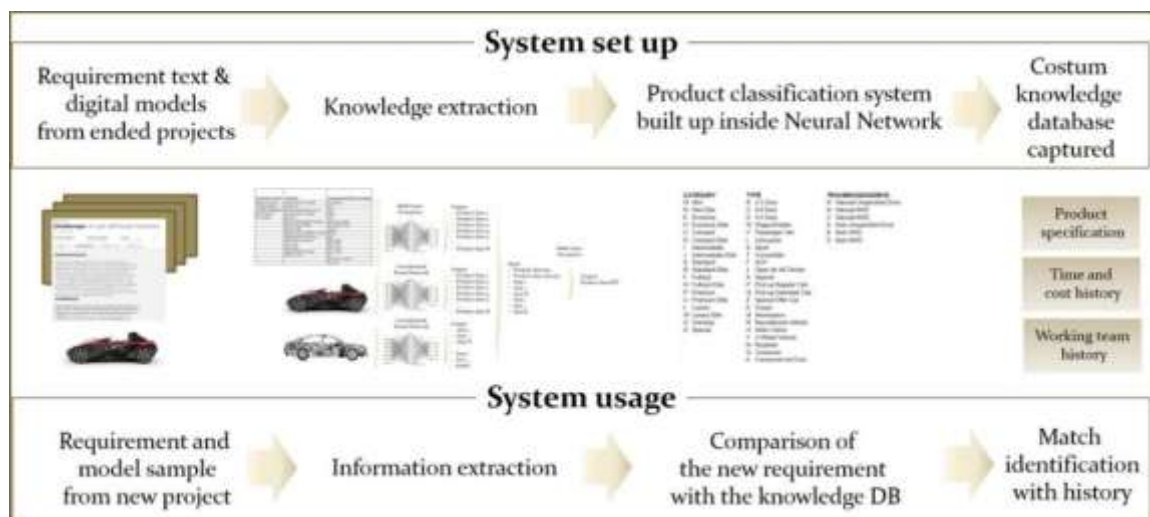


Figure 3: InnoCrowd overview, system set up and system usage

The block "System usage" shows the way we use the DSS to decide on a new project. Requirements and model samples are input to the system. The second step is the extraction of relevant information from the given dataset (input in the previous step). In the third step, the DLNN classifies the new project by comparing it with others in the knowledgebase. The final step provides project-specific information relevant to a decision: product specification, related time and cost, and working team requirements. The user can then decide whether to apply for the new project.

We use DLNN for the automated recognition of the product design. For each step there is a special neural network trained for recognition. A deep learning neural network is known as the best analyzer / classifier of 2D pictures. It can process the data faster than a 3D model classifier [Sarvepalli, 2015] .

The requirements are collected as free form text from the source system (crowdsourcing provider platform). Each relevant part of the text will be mapped to the related domain or application or adaptation layer. Then the Neural Network classifies the proposed product.

1.5 Dissertation Contents

The thesis is structured as follows. Chapter 1 introduces the topic, explains the work scope, and discusses the contribution. Chapter 2 defines relevant processes, methods and themes. Chapter 3 explains crowdsourcing problems in industry. Chapter 4 reviews the state of the art in crowdsourcing. The first, second, third and fourth chapters are bundled as a first part of this thesis and represent the background and problem statement.

The second part explains InnoCrowd, the system we developed. In the fifth Chapter we explain methods for analyzing the text and image data for building a knowledge. Chapter 6 shows the optimization of the algorithm for data classification.

Finally the third part shows how the method is applied in three industrial use cases, namely for engineering product development, for architectural product design and also for software as product created by the crowd (Chapter 8). Chapter 9 validates hypothesis. Chapter 10 summarizes and provides suggestions for future work.

Chapter 2 BACKGROUND AND DEFINITIONS

2.1 The Creation of Innovative Products

In the past, company product development departments were often closed units. Even today, the exchange of information between departments in the company is sometimes characterized by battles for resources, and the exchange with external partners is often characterized by the fear of losing control over valuable knowledge. However, organizations are increasingly realizing that valuable, innovation-relevant knowledge can also be found by connecting with customers, employees and other stakeholders. By including these other parties in previously internal, closed processes, distributed knowledge can contribute to the success of new products and services. Consumer goods manufacturers who use crowdsourcing for new product development may benefit from a 1% increase in sales [Gartner, 2014].



Figure 4: Traditional vs crowdsourced task completion (source: Innosabi)

The advantages of co-creation are also reflected in the economic success of products developed through crowdsourcing--higher turnover and higher margins compared to exclusively in-house developed products. In addition, companies that open their product

development through crowdsourcing can position themselves as innovative, transparent and customer-oriented and thus enhance their brand. Because clear changes are also noticeable on the part of customers: They know about their power in a buyer's market and the importance of their purchase decisions. In addition, through social media, the opportunities to make yourself and your wishes heard and reach have increased rapidly. So it is not surprising that crowdsourcing has been gaining in importance in product development for several years.

In the current “Age of the Customer” before users get in contact with a producer to receive further information about material or a product sample, they will find out about the company’s products and services [Schade, 2019].

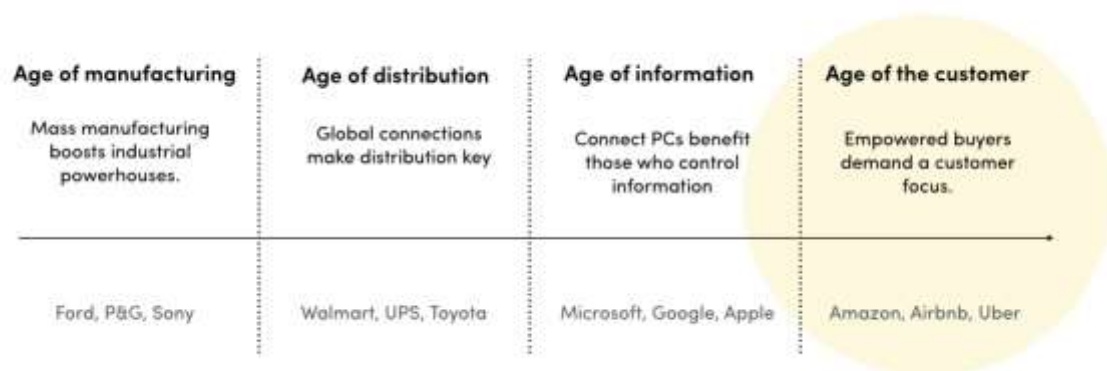


Figure 5: From the age of manufacturing to the age of the customer (source: medium.com)

Customers have become powerful. Digital technologies and innovations in the last ten years enable customers to make complex purchasing decisions independently. The approach taken by companies to focus first on their own brand and only then on the interests and needs of their customers can no longer work. Crowdsourcing is an effective approach for adapting to this new age.

2.2 Overview on Crowdsourcing

2.2.1 Definition of Crowdsourcing

The term crowdsourcing was defined by Howe by 2006 [Howe, 2006] to represent a wide group of activities that take on different forms.

“Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call”.

Nowadays, 13 years after the first definition, crowdsourcing has gained popularity in several application areas because of significant cost and time savings [Eagle, 2009]. It applies to many stages of product development, including simple tasks such as tagging a picture or programming a simple app, and complex tasks like design and production of a vehicle [Local Motors, 2004]. Crowdsourcing can be effectively used by companies for task delegation. Crowdsourcing is an effective method for getting expertise when it is not available internally [Chatterjee, 2014]. For example, the definition of product requirements and getting feedback on product performance are areas where input from external sources is more relevant, because the crowd can represent potential customers better than company employees can [Borchert, 2012]. Still, it is so difficult to determine the risk of complex projects and to decide whether to use crowdsourcing that managers need special methods and tools [Thuan, 2015].

2.2.2 Industrial application of crowdsourcing

Early applications of crowdsourcing handled simple tasks such as text translation and picture tagging. Later, people extended the use of crowdsourcing into product design. Now it also covers product manufacturing. Wu proposed the new term “Cloud Based Design and Manufacturing” (CBDM) to represent this development [Zheng, 2016]. CBDM refers to a service-oriented networked product development model in which service consumers are enabled to configure, select, and utilize customized product realization resources and services ranging from computer-aided engineering software to reconfigurable manufacturing systems. Gibb called his similar approach “Open-Source Hardware” [Gibb, 2015] [Prpic, 2015]. Design

of open-source hardware is hardware whose design is made publicly available so that anyone can study, modify, distribute, make, and sell the design or hardware based on that design. The hardware's source and design are available in a convenient format for making modifications to it. Ideally, open-source hardware uses readily available components and materials, standard processes, open infrastructure, unrestricted content, and open-source design tools to maximize the ability of individuals to make and use hardware. A good example for the realization of “Open-Source Hardware” is the car development and manufacturing by the company Local Motors. Local Motors is supported by more than 30.000 engineers (the crowd) for executing the design process of a new car. Traditional automobile manufacturer employs 1000 to 5000 engineers to execute almost the same task. Having a greater number of people engaged during the engineering process accelerates the design process 300% and makes the manufacturing cost more than 10 time cheaper [Eagle, 2009].



Figure 6: Crowdsourced vehicle development in Local Motors

The design phase produces a 3D model of the car. Each sub process produces its own product in the form of information which has to be managed, including files generated from design, engineering and manufacturing, such as 2D drawings, 3D models, machining files, etc. Metadata includes information such as product structure, requirement documentation, competition project specification, information on crowd members, project team, voting, etc. Amazon Mechanical Turk, one of the best-known crowdsourcing platforms, offers micro tasks to the crowd, for example tagging images, and translating fragments of text. Top Coder offers low complex software tasks that together respond to user requirements. Similarly, Cad Crowd

divides up new product design into smaller jobs. We used data from Cad Crowd applications to train our machine learning algorithm and for testing.

2.2.3 The classification of crowdsourcing

Martin [Martin, 2008] created a matrix for the classification of crowdsourcing. The aspects "not innovative", "innovative", "amateur" and "specialist" are associated with another group of aspects about the knowledge and experience that actors need in order to cope with the given tasks. Amateurs can handle some of the simpler tasks. Martin includes the subject areas of open innovation and open source in their classification, as Figure 7 shows.

Quadrant I represents platforms that need only highly qualified specialists for the development process of innovative products or services, such as the InnoCentive and Nine Sigma platforms. Platforms in the second quadrant are a further development of the original outsourcing. The task areas have a non-innovative character but require handling by specialists. The main focus is on finding new trends and developments. Trend watching is a well-known platform where journalists report on innovative trends. Another example is the Marketocracy platform. Skilled investors and their strategies are sought there. No innovative know-how is required in the third quadrant either. However, this quadrant is characterized by the inclusion of amateurs and / or hobbyists. Above all, the creative area of crowdsourcing is addressed, for example on platforms such as Threadless or iStockPhoto.

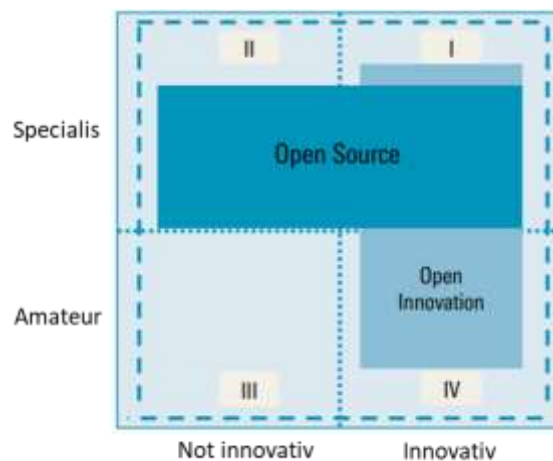


Figure 7: Crowdsourcing classification

However, the transitions between Quadrant II and III are partly blurred, and a differentiation of crowdsourcing providers and crowd work projects appears difficult. Platforms that have micro jobs can be assigned to the third quadrant, for example, Mechanical Turk from Amazon, which is also very close to classic outsourcing. Due to its close proximity to outsourcing, Quadrant III has the highest relevance for crowdsourcing. In the fourth and last quadrant, the authors do not currently see any platforms, since in innovative processes participants without special knowledge are usually involved only in the adaptation of a product to improve customer service. Thanks to the four fields and the high degree of selectivity, platforms can be assigned more precisely than with the previous variants. However, the “non-occupation” of the fourth quadrant could be a problem, because crowdsourcing should allow amateurs to participate when appropriate. A large number of hobbyists have broad and in-depth knowledge in their areas without formal education or training.

2.2.4 Crowdsourcing process as an innovation platform

Crowdsourcing as an innovation platform is characterized by clear focus on innovation development. The process starts with the definition search field, where everybody gives any suggestion on new innovation topic and idea (see Figure 8). The second step is when the scope is already defined. The crowd starts to generate ideas focused on the defined scope. The third step is the prioritization of some winning ideas. The fourth step is the development of a product based on the prioritized product ideas. The fifth step is market introduction of the final product from the crowdsourcing process.

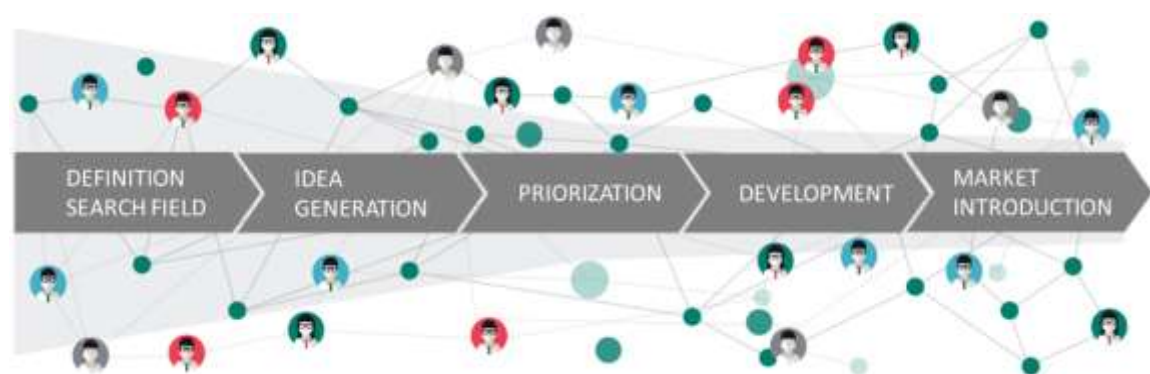


Figure 8: Crowdsourcing process (source: Innosabi)

For the majority of the projects carried out via these platforms, crowd creation is also carried out in addition to crowd voting, i.e. the solutions submitted are publicly available and can be rated and supplemented by the other members [Yang, 2016]. Crowdsourcing sometimes compensates with non-cash prizes. With 55,000 members, Jovoto is the largest innovation platform in German-speaking countries. Since entering the market in 2008, more than 250 projects have been carried out with customers such as Coca Cola, Starbucks and Deutsche Telekom in the field of creative solutions and generating ideas. The cooperation between the members plays an important role in the innovation competitions. In this way, the members can comment on and rate the ideas of others and receive points for this. In contrast to competitions in the “Design” cluster, there is usually no individual winner here. Typically, the client pays between two and 25 crowd workers at a time. The premiums for first place winners are usually between 1,000 and 3,500 euros. The UnserAller platform is somewhat smaller with a community size of around 20,000 co-developers. It has been used since 2011. Around 25 projects have been completed so far, each lasting several months. The client does not pre-select the crowd workers. All members can contribute to the process in the individual phases by posting suggestions and comments, voting or developing prototypes. Several participants are awarded in the form of prizes in kind (e.g. vouchers) that are directly related to the customer's range of goods. Large platforms like Jovoto and Aitzo have membership numbers of tens of thousands of crowd workers. eYeka, a crowdsourcing platform founded in Paris, has even more than 150.000 member. Such intermediaries have already handled several hundred projects and given crowd workers financial incentives to participate. Numerous smaller platforms and platforms with a much lower project throughput rate have emerged, such as UnserAller and bonspin.

2.3 Management of Crowdsourcing

Hetmank (11) defined 4 main aspects to be managed in crowdsourcing: user, task, contribution and workflow (Figure 9). User management is comparable with the organization module of an engineering product development system. It defines detail on user, roles, working group, etc. Task management is comparable with management of the manufacturing process in an engineering product development system. It covers the job sequence and the assignment to a certain stage and user. Contribution management is the management of results produced by

the crowd. In an engineering context these results are 3D models. 3D models and workflow management are standard modules in an engineering product development system.

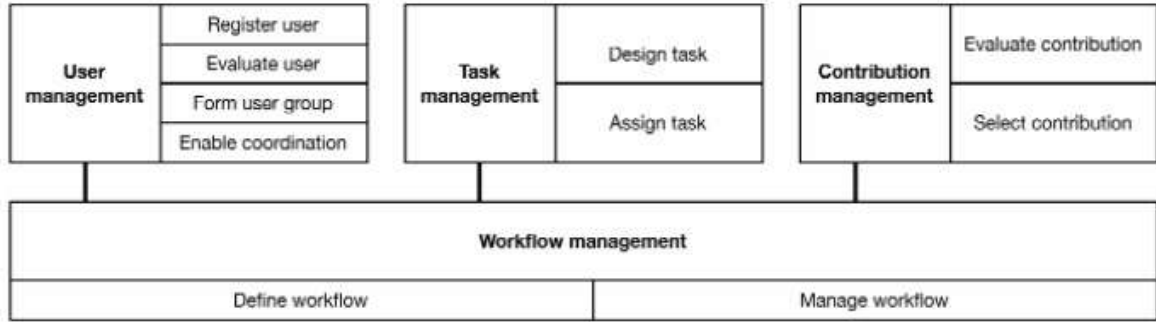


Figure 9: Main aspects to be managed in crowdsourcing

We found that two more important aspects than Hetmank's: crowdsourcability and platform. Platform refers to the state of the art in technical implementation of crowdsourcing tools.

2.3.1 Crowd management in context of crowdsourcing

Kim et al. suggest the possibilities that arise if we start to think of crowdwork not just as a collection of tasks to complete but as a collaborative activity that workers themselves can influence. Wisdom—even that of the crowds— comes not from blindly following orders but from dialogue, reflective practice, and revision. Malone et al. [Malone, 2011] developed a system which suggests that crowd-workers can either work alone independently or depend upon each other to work together. Zhao [Zhao, 2015] developed a model that provides a principle and natural framework for capturing the latent skills of workers as well as the latent categories of crowdsourced tasks. The inference of the latent skills of workers is based on past resolved crowdsourced tasks with feedback scores. Mrass [Mrass, 2012] developed methods for crowd assessment that ensure the required qualifications of the (often unknown) crowdworkers, since education and skills are an important way to tackle complexity. Haas [Haas, 2014] developed Argonaut, a system that uses a predictive model of worker quality to select trusted workers to perform review, and a separate predictive model of task quality to decide which tasks to review. Argonaut can identify the ideal trade of between a single phase of review and multiple phases of review given a constrained review budget in order to maximize overall output quality. Valentine [Valentine 2014] developed a system that organizes crowd workers into computationally-represented structures inspired by those used in

organizations — roles, teams, and hierarchies—which support emergent and adaptive coordination toward open-ended goals. It introduces two technical contributions: 1) encoding the crowd’s division of labor into de-individualized roles, much as movie crews or disaster response teams use roles to support coordination between on-demand workers who have not worked together before; and 2) reconfiguring these structures through a model inspired by version control, enabling continuous adaptation of the work and the division of labor. Cui [Cui, 2020] proposed incorporating a policy network for the selection of task allocation strategies and a reputation network for calculating the trends of worker reputation fluctuations. The research results explained above can be summarized as follow:

- Malone et al [Malone, 2011] Manage crowd collaboration
- Zhao [Zhao, 2015] Manage crowd skill
- Valentine [Valentine 2014] Manage crowd structure
- Cui [Cui, 2020] Manage crowd reputation

The crowd should be managed as an organization, similar with the approach from Valentine. The crowd can be handled as an object in a data management system. It can be revised and thrown into a certain workflow for getting a new status (crowd promotion process).

2.3.2 Task management

Dunhui Yu [Yu, 2019] models the collaborative software task assignment problem in the crowdsourced environment as the assignment optimization problem and integrates the three factors of worker capacity, task module complexity and worker active time to establish optimization goals. Tran-Thanh [Thanh, 2015] developed CrowdForge, a framework that provide a systematic and dynamic way to break down tasks into subtasks and manage the flow and dependencies between them. Tran-Thanh developed an algorithm to first calculates an efficient way to allocate budget to each workflow. It then determines the number of inter-dependent micro-tasks and the price to pay for each task within each workflow, given the corresponding budget constraints. Aipe [Aipe, 2016] investigated the role of task similarity in microtask chains and how it affects worker performance. They model overall task similarity of a task pair and studied the impact of task similarity in microtask chains on worker retention, satisfaction, boredom and fatigue. Their studies reveal that ordering tasks in a chain according

to overall task similarity results in improved accuracy, but at the cost of inducing boredom. Kittur et al [Kittur, 2011] developed CrowdForge, a system that explains how map-reduce framework popularized by Google may be used to partition bigger complex tasks into smaller tasks dynamically by workers. Rodriguez [Rodriguez, 2014] applied an automated classification system for supporting task management. The system classifies images automatically. The crowd must validate the result and give the correct annotation to each image. This allows a faster annotation of the classified images. This approach becomes complex when the object to be classified is not a single object but has a complex structure such as image of a car. Our work addresses the handling of a complex structure, using a car as an example. Schmitz [Schmitz, 2016] developed TAS, a model that adds the timeline and online perspective to task assignment optimization in expert crowdsourcing. André [André, 2017] stated that a sequential work structure was more effective than a simultaneous work structure as the size of the group increased. The research results explained above can be summarized as follow:

- Dunhui Yu [Yu, 2019] Task assignment optimization: task complexity, time
- Tran-Thanh [Thanh, 2015] Task break down & budget consideration
- Aipe [Aipe, 2016] Task similarity & performance
- Rodriguez [Rodriguez, 2014] IT tool for partial task execution
- Schmitz [Schmitz, 2016] Timeline and online task assignment optimization
- André [André, 2017] Sequential vs simultaneous work structure

The task can be seen as a requirement. It is a product requirement in context of engineering product development. If we have a history (knowledge database) of the relation between the requirement and the product then we associate the new task (requirement) with the old task already executed. This approach is to a certain extent related to the approach from Aipe. We need to establish the knowledge database to represent the relation between the requirement and the product.

2.3.3 Contribution Management

Hirth [Hirth, 2012] showed that crowd-based cheat-detection mechanisms are reliable, easy to implement, and cheaper than manual processes. Sautter [Sautter, 2011] developed mechanisms that capitalize on especially capable users and enforce data quality with less decrease in

throughput than occurs with static, redundancy-based approaches. Allahbakhsh [Allahbakhsh, 2013] proposed a framework for characterizing various dimensions of quality control in crowdsourcing systems.

Research results can be summarized as follows:

- Hirth [Hirth, 2012] Cheat detection mechanism
- Sautter [Sautter, 2011] Task routing to capable users
- Allahbakhsh [Allahbakhsh, 2013] Manage critical issues

Contribution management in the context of engineering product development focuses more on data management than issues like cheat detection. It deals with 3D models that the users (crowd) share to improve their collaboration. In the context of crowdsourcing, 3D models are saved in a single database system. Standard data management methods such as search engine, revisioning and release status are not applied yet in the crowdsourcing context.

2.3.4 Process / workflow management

Little et al. [Little, 2009] developed TurKit, a system that helps decide what to present to each worker such that the flow of results of tasks between dependent workers can be controlled. Shilovitsky [Shilovitsky, 2012] reported that data produced during the crowdsourcing process is not well managed, and that processes are not clearly tracked in crowdsourcing platforms. Standardized workflows need to be defined and applied. Yuen [Yuen, 2011] developed a strategy for matching tasks with the profile of each worker by using the past task preference and performance of each worker to determine when the worker is ready to work on the next task.

This paper also explores reusing past information. Geiger et al. [Geiger, 2012] constructed a recommendation system matching individual in the crowd with types of crowdsourcing tasks. Vaish [Vaish, 2013] developed a framework to allow users to systematically apply operators such as split and merge on project ideas, code, or crowd. The research results can be summarized as follow:

- Little et al. [Little, 2009] Task result flow between dependent workers
- Shilovitsky [Shilovitsky, 2012] Workflow for data management
- Yuen [Yuen, 2011] Matching tasks with the profile of each worker.

- Vaish [Vaisch, 2013] Organizing the process, ideas, code or crowd

Routing the task to the worker should be made easier. The worker can be attached to the structure of a task. The concept resembles the management of manufacturing processes, where process structure and operator are attached to the related manufacturing sub process. A history of relationships between task and worker can also be established in order to enable reuse [Kusumah, 2020]. We can apply to workflow the best practices from the engineering development process.

2.4 Natural Language Processing for Text Analysis

Natural Language Processing (NLP) automatically recognizes human language. NLP can be broken down into Natural Language Understanding (NLU) and Natural Language Generation [Verspoor, 2013]. The former is used understand text and language (semantic connections and the context of the content), while the latter relates to the creation of text or spoken narratives from structured data.

The most recent developed concepts of NLP are word embeddings, language modeling and transformer models.

2.4.1 Word embeddings

Word embeddings works by assigning words numerical values, so vectors stand for similar concepts, for example “king / queen” or “man / woman”. In 2018, this followed the principle of a dictionary with rules and synonyms. That opened up a world of opportunity for NLP. For the first time, as words could be processed in terms of their importance. Classic word embeddings are a challenge. For example, it is hard to interpret the meaning of ambiguous words in different contexts.

The best-known word embedding models (such as glove and word2vec) are based on matrix factorization [Li, 2017]. For a given vocabulary of size n , an input matrix of size $n \times n$ is created from training data, which records relationships between words in the entire text corpus. The input matrix counts how often a word occurs in connection with other words (for example in the same sentence or in a specified environment); this is also called co-occurrence. This matrix

is then factored into matrices of smaller dimensionality so that the information from the input matrix is compressed into smaller dimensions. This means that relationships between words must flow into the smaller representations. Thus, the meaning of a word is defined from the words that often occur in its environment - matching the well-known quote from the well-known linguist John Rupert Firth: [Firth, 1957] "You shall know a word by the company it keeps".

Although word embeddings are very useful for many tasks, they also have significant disadvantages: In order to represent all of the information in the training data, the vectors must have a certain size. Common models generate vectors between 50 and 300 numbers in length. This means that a word can be represented by up to 300 numbers. It is not clear whether individual numbers represent different properties of the word or whether only the geometric relationships of the complete vectors determine the representation. Due to the lack of traceability, it is hardly possible to examine a model that classifies text on the basis of these vectors in more detail. If a word is classified incorrectly, the error can possibly be traced back to individual dimensions of the input data (word vectors). However, the individual dimensions cannot be easily interpreted, which makes it difficult to continue the analysis.

2.4.2 Language modeling

Language modeling is the next level. A classic task of this process is to begin with an analysis of a large corpus (text) and then it will predict the context of the words. It will recognize syntax, context and also the specific domain. The extraction of word embeddings from Language Models brought thus a further semantic advance: for every meaning of a word a new vector is assigned in the best case [Qudar, 2020].

This principle has been known as Masked Language Modeling (MLM) and the underlying system is called BERT. BERT was trained to fill in "masked" word gaps correctly in which there are preceding and following words. BERT is thus able to do the fill the following example with meaningful terms [Chang, 2018]. This also shows the possibility that BERT is able to translate sentences through their related context, which is called Next Sentence Prediction (NSP). Through these facts BERT shows new possibilities, especially that the algorithm will be able to understand context better, which makes NLP better in the future and can offer communication opportunities.

2.4.3 Transformer models

Transformer models enable algorithms to be trained in an attention-based manner. It helps to recognize and focus attention of the most important part of each sentence [Wang, 2019]. This approach helps to recognize the relationship between words that are not next to each other. To do this, transformers break down sentences in different ways to get information about each word's context. This reduces the training effort for algorithms and speeds up use of large data sets.

In summary, all concepts, starting with word embeddings, strive for training efficiency requiring as little domain-specific data as possible. Also they try to model semantics effectively.

2.4.4 Related application

The potential of NLP lies in the automation of language-based procedures and processes. It can directly access natural language information without requiring labor to encode it. Classification can be used for intelligent filtering of text content and documents (e.g. e-mails) or for purposes of routing. For example, inquiries can be sent to the responsible contact person and experts within an organization. Other areas of application include automated extraction, summary and categorization using topic modeling in order to analyze the content of for example news, reports, contracts or social media content. Another area of application is sentiment analysis, which records moods in texts, to analyze whether customers are familiar with the products and services of the company. These are performed by AI-based dialogue systems, such as Chatbots, which also use NLP technology (mainly NLU) and get more and more widespread distribution. NLP can create added value in almost all corporate functions. In particular for recruitment, marketing, customer service, the internal process control or the market observation. Also in healthcare (for analysis of the cause of illness) or in the administration (for example in the Federal Office for Migration and Refugees) there are numerous possible uses for NLP.

The difficulties for NLP are: language does not always follow strict logical rules. It is dominated by emotions and changes frequently depending on the situation in which it is used. It is extremely difficult for an algorithm to detect sarcasm, irony, or hidden criticism, for example. The linguistic data are unstructured - nevertheless they form the basis for programs that work

exclusively according to logical rules. However, according to the principle of trial and error, software gradually recognizes special contexts and thus avoids misinterpretations - for example in translations. The more standardized and structured the data, the easier it is for digital systems to process the data. Human language is inherently complex. The rules of human communication are rather coarse and loosely structured. This is where NLP comes in and tries to recognize these structures.

Chapter 3 PROBLEM STATEMENT

The manager of a new crowdsourced project must vet ideas, primarily takes into account their suitability for the brand and company as well as the desired level of innovation and feasibility within constraints such as time and budget. In the first step, companies should define their objective, i.e. what added value they want to achieve by working with customers. Does the company want to jointly develop a completely new type of product that reinvents an entire product category, or is their goal to quickly-extend their product line with tight framework conditions? Should ideas primarily be generated from the customer experience, or is it about identifying user needs that are still unknown through the discussion? Crowd approaches for generating ideas and solving problems usually require the involvement of product management and the research and development department in communications in social media with those responsible for marketing and communication.

In the next step, the interdisciplinary project team formed in this way should define the framework conditions for the product to be generated: If a high level of innovation is required, these are usually based on the overall corporate goals or overarching issues, e.g., “how will demographic change affect demand for our product?” If the level of innovation is rather low, the production and implementation options are the limiting factor.

In order to allow the project to proceed in a targeted manner and to complete it with the greatest possible success, a detailed project questionnaire is essential. In addition, the manager should appoint for community management and moderation of the course of the project.

The sub chapter below outlines research into supporting managers lacking crowdsourcing experience.

3.1 Crowdsourcability

Afuah [Afuah, 2017] stated that every task is crowdsourcable as long as it can be easily described, communicated, and modularized. Afuah described four factors a manager must

handle when their company wants to use crowdsourcing: task characteristics, the availability of the crowd, cost, and infrastructure. Knop [Knop, 2019] stated that a clear and well-protected IP may be important for the decision to crowdsource. Predictable costs and stable requirements are necessary for crowdsourcability. Ford [Ford, 2015] stated that the benefits and quality of crowdsourcing are not enough to convince a company if costs would increase. Yuen et al. 6 [Yuen, 2011] and Prokesch [Prokesch, 2014] proposed decision support for processing outcomes from the crowd. Thuan built an enterprise ontology of business process crowdsourcing. It includes the main business processes, data entities, data attributes, and their hierarchy relationships, which were structured into a lightweight ontology. He added decision-making relationships and business rules that turn a lightweight into a heavyweight ontology. These research results can be summarized as follows:

- Afuah [Afuah, 2017] Task modulation, crowd availability, cost and infrastructure
- Knop [Knop, 2019] A clear and well-protected IP
- Ford [Ford, 2015] Balance of benefits and quality
- Yuen [Yuen, 2011] Decision support for processing outcomes from the crowd
- Thuan [Thuan, 2015] Decision-making relationships and business rules

Research has confirmed that task modulation defines the grade of crowdsourcability. The nature of product requirements determines the feasibility of crowdsourcing. Thuan's method may not be appropriate for crowdsourcing tasks because it is not specific enough. We are developing a decision support system for crowdsourcing based on information extracted from each vendor, so that the calculation of crowdsourcability of new crowdsourcing project can be done precisely, because it is based on internal information about the crowdsourcing vendor [Kusumah, 2021].

3.2 Platform

The crowdsourcing platform of Local Motors consists of two main tools, namely a 3D design tool and a website. The website is used for managing the competition / challenges, discussion forum, storing files and project management. TopCoder utilizes several collaboration tools in their platform [TopCoder, 2015]. GitHub and GitLab are used for code and project

management. Trello manages requirements. A Chrome browser extension manages issues and tasks. TopCoder provides a UML tool for use in development competitions. It helps model processes, classes, use cases, and activity diagrams. Ebert [Ebert, 2014] compared several platforms for crowdsourcing support based on these criteria: communication, collaboration and coordination (Table 1).

Platform	Communication	Collaboration	Coordination
Crowdsourcing platform	Fair	Poor	Fair
Platform as a service (PaaS)	Poor	Fair	Fair
Open source platform	Good	Good	Fair
Collaborative testing platform	Good	Poor	Fair
Enterprise collaboration platform	Good	Good	Good

Table 1: Crowdsourcing support from various software development platforms.

Bonvoisin [Bonvoisin, 2014] conducted a comprehensive study of existing platforms. Figure 10 shows tools for tasks. Because transfer of data among tasks is mainly done by manual data export and import, information could be lost or corrupted. Time-consuming conversion into a neutral file format is required before transferring data to the next tool.

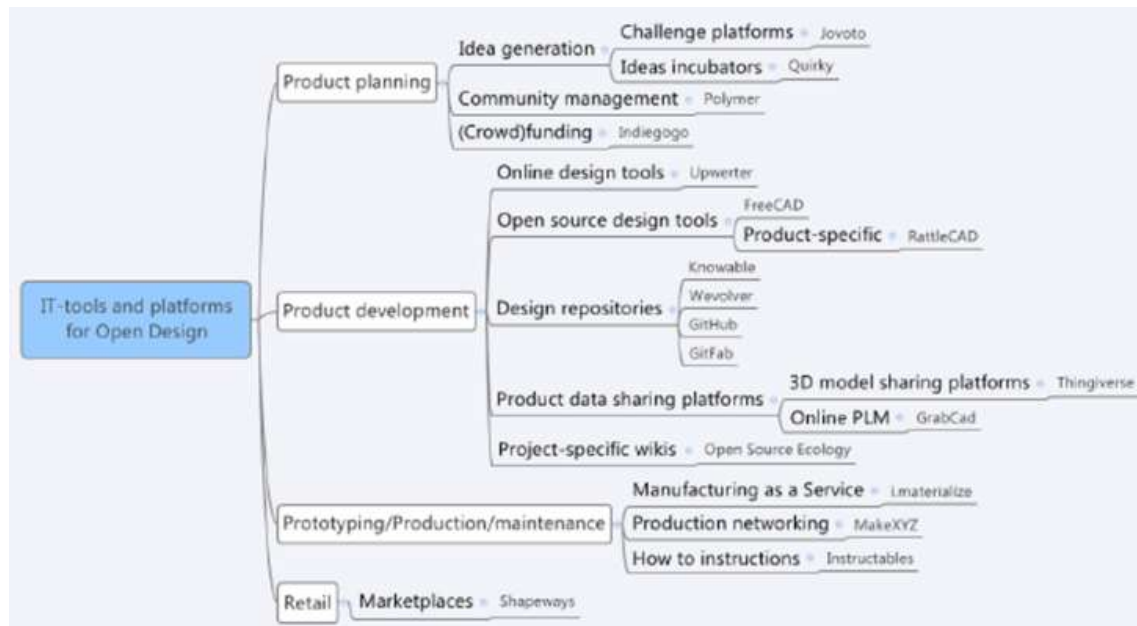


Figure 10: IT tools and platforms for open design

The platforms and research results explained above can be summarized as follows:

- Local Motors [Local Motors, 2004] 3D design tool and a website
- TopCoder [TopCoder, 2015] GitHub, GitLab, Trello
- Ebert [Ebert, 2014] Communication and Collaboration
- Bonvoisin [Bonvoisin, 2014] Tools for product planning & development

A primary shortcoming of existing crowdsourcing applications is that they use “traditional” approaches for managing processes and data. Folder-based data management makes the information not easily accessible and wastes time. Files are accessible for all users, but there is no versioning feature for the design object, so that nobody can tell exactly the actual version of the current development. These tools are stand alone. They are good for handling semi-complex tasks, but not sufficient for handling large product structures common in car development. A system lacking a versioning feature is not appropriate for collaboration. Everyone should be able to find the right and actual design object for collaboration involving different people from different places. An enterprise collaboration platform can sufficiently handle the three aspects, but it is not applicable for crowdsourcing in the automotive industry

because it doesn't handle CAD files. Typically, managers choose open-source platforms for software production. Platforms have detailed features for communicating software requirements and for code management, but generally have no feature for crowd management. These aspects that are not sufficiently fulfilled by the current crowdsourcing platform make the decision to crowdsource harder.

3.3 Gaps Along the Crowdsourcing Process

Nevo [Nevo, 2020] identified 4 gaps in knowledge of how to manage crowdsourcing.

- Gap 1. Crowdsourcing Tends to Have a Narrow Perspective. Research is needed to determine how to make the decision to crowdsource using a broader, organizational perspective, so that crowdsourcing is considered along with other approaches it is expected to complement.
- Gap 2. How to Evaluate the Success and Value of Crowdsourcing Initiatives? Research is necessary to explain different approaches appropriate to handle one-off crowdsourcing projects versus ongoing crowdsourcing initiatives. The latter are likely to be designed as long-term strategic initiatives aimed at delivering new product ideas.
- Gap 3. Horizontal Integration: The Role of Platforms in the IS Crowdsourcing Literature. It is necessary to perform a longitudinal examination of the role of platforms throughout the lifecycle of a project, with attention given to interactions among stakeholders in each stage.
- Gap 4. Vertical Integration: Synthesizing Micro-level Insights Using Specific Lenses. We need to interpret what knowledge can be gained by analyzing connections between micro-level studies. Also beneficial would be deeper studies (e.g., multi-level firm-centered studies) and broader ones (e.g., large crowd-centered, platform-centered or industry-centered studies).

All these gaps identified by Nevo show immaturity of crowdsourcing at the process level.

3.4 Research Hypothesis

Based on the issues and gaps discussed in the previous sub chapter we formulate our research hypothesis as follow:

1. Best practices for performing a crowdsourced product development project are available but not commonly known by the new user / crowdsourcer
2. There are only limited methods available for handling the problem in a comprehensive way
3. Crowdsourcing project risk can be made transparent qualitatively and quantitatively
4. Artificial Intelligence can deliver a qualified performance for data mining of the complex crowdsourcing project

The hypothesis above also reflects the sequence of this PhD work, where we started with literature review in Chapter 2 in order to search for relevant best practices. And more important is to find whether the best practices are known and applied by the industry.

The discussion of the second hypothesis will differentiate between methods that handle problem in very specific aspect of crowdsourcing and methods that handle the crowdsourcing in a comprehensive way. It is whether to go deep in specific direction or to handle the most important aspects in parallel. The second approach is more related to the implementation of crowdsourcing in the industry, which belongs to the scope of this work. This aspect will be discussed in the Chapter 3 where we give the problem statement.

The third hypothesis will be discussed in the fourth Chapter where we explore the state of the art on product classification. The methods discussed in that chapter will be the base for building a process chain to make a crowdsourcing project risk become transparent. The process chain represents the engine of InnoCrowd, the classification tool for a decision support system in a crowdsourced product development.

Chapter 5, 6 and 7 will give an answer whether artificial intelligent can deliver a qualified performance or not for the mentioned challenge (the fourth hypothesis).

Chapter 4 STATE OF THE ART ON PRODUCT CLASSIFICATION

The second Chapter and the third Chapter have discussed the first and the second hypotheses. In this fourth Chapter the focus is on state of the art for discussing methods that are usable for giving an answer to the third hypothesis: “Crowdsourcing project risk can be made transparent qualitatively and quantitatively”. The answer is related directly with the explanation of InnoCrowd in the first Chapter. Figure below gives an overview of the answer (InnoCrowd concept). At the end of the process chain of InnoCrowd it is to be made transparent, namely the relevant information that makes the crowdsourcing project risk become clear qualitatively and quantitatively to the new user. This information are product specification, time & cost history and working team history.

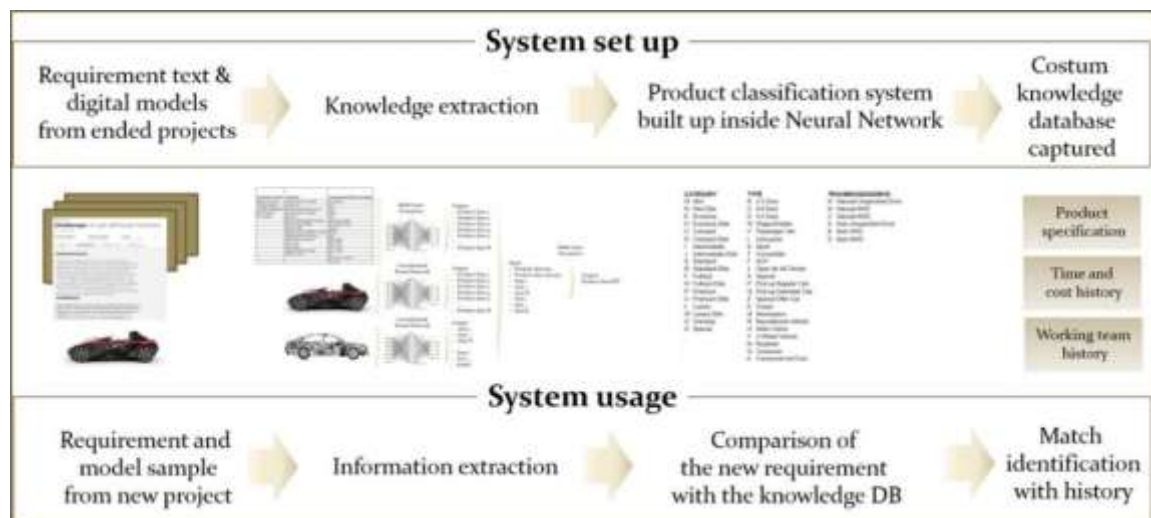


Figure 11: InnoCrowd overview, system set up and system usage

In the middle of the process chain InnoCrowd needs functionalities / methods to be used for product classification. Therefore, the state of the art discuss this topic and it serves as starting point for the development of InnoCrowd.

4.1 Product Classification Procedures

The main task of InnoCrowd is the classification of a product based on its requirements and images.

Classification procedures can best be classified on the basis of these properties [Wikipedia]:

- Manual and automatic procedures
- Statistical and non-distribution methods
- Supervised and unsupervised procedures
- Parametric and non-parametric methods

4.1.1 Manual and automatic procedures.

Automatic methods employ a formal method for making decisions in new situations on the basis of learned structures. These methods generate an algorithm (the learning algorithm) and then apply it to known and already classified cases in a database to calculates structures [Figuerola, 2001]. These newly learned structures enable another algorithm (the evaluating algorithm) to assign a new and previously unknown case to one of the known target classes on the basis of the observed attributes.

4.1.2 Statistical and non-distribution methods.

Statistical methods are based on density calculations and probabilities, while non-distribution methods use clear dividing surfaces to separate the classes [Bakir, 2001]. The boundaries between the individual classes in the feature space can be specified by a discriminant function. Examples of statistical methods are the Bayesian classifier, the fuzzy pattern classifier and the kernel density estimator. The calculation of parting areas is possible using support vector machines.

4.1.3 Supervised and unsupervised procedures.

The creation of structures from existing data is also known as pattern recognition, discrimination or supervised learning [Ciocca, 2006]. Classifications are given in the dataset. In unsupervised learning, the classes of the data are not specified but must be learned. An example of unsupervised procedures is cluster analysis. In reinforcement learning, information can be added about whether a class division was correct or incorrect.

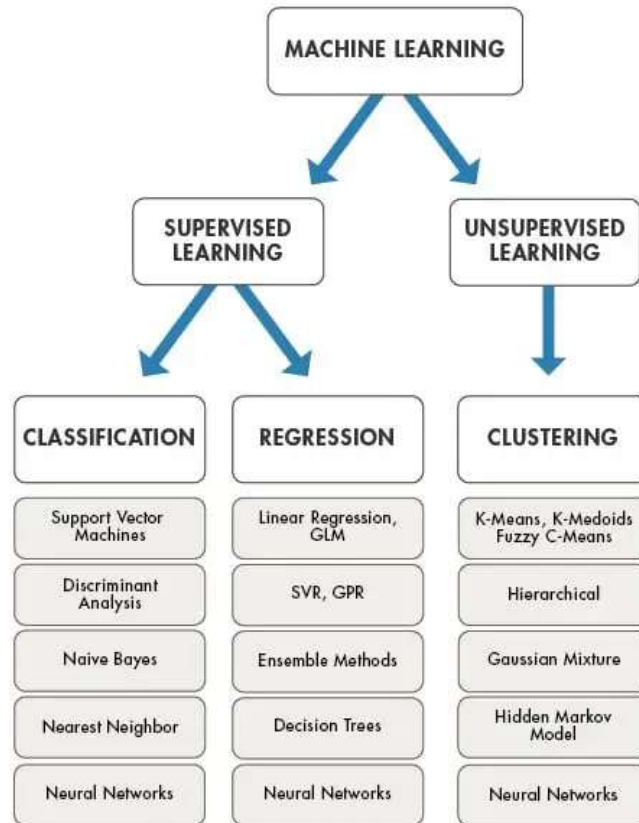


Figure 12: Supervised and unsupervised procedures.

The ability of intelligent algorithms to learn goes far beyond the training phase. Thanks to supervised machine learning, algorithms are able to improve rules that have been learned once they receive the appropriate feedback. In addition, once they have learned rules, they can apply them to new cases or unstructured data. Once an algorithm has been trained to understand the spoken language, it is possible to subsequently adapt it so that it can also understand one or more dialects. Large amounts of data are advantageous when it comes to finding use cases in

companies for learning algorithms. Not least for this reason, supervised machine learning has become one of the most important methods in machine learning.

4.1.4 Parametric and non-parametric methods

Parametric methods are based on parametric probability densities, while nonparametric methods (e.g. nearest-neighbor classification) are based on local density calculations [Kumar, 2012].

4.2 Supervised Procedures for Product Classification

Machine learning plays a central role in the development of software and algorithms for image recognition. Teaching adaptive programs about what can be seen in pictures is quite challenging. During the learning process, the algorithms have to compare millions of images with one another in order to identify similar patterns in the images [Rucco, 2018]. It is a complex and resource-intensive process. The following sub chapters explain the most important algorithms of supervised procedures for the classification task.

4.2.1 Convolutional Neural Networks (CNN)

A convolutional neural network is a deep learning architecture that was specially developed for processing images [Bezdan, 2019]. In the meantime, however, it has been found that convolutional neural networks also work extremely well in many other areas, e.g. in the area of word processing. It is able to process input in the form of a matrix. This enables images displayed as a matrix (width x height x color channels) to be used as input. A normal neural network, e.g. in the form of a multi-layer perceptron (MLP), on the other hand, requires a vector as input, i.e. in order to use an image as input, the pixels of the image would have to be rolled out one after the other in a long chain (flattening). As a result, normal neural networks are e.g. not able to recognize objects in an image regardless of the position of the object in the image. The same object in different position in the image would have a completely different input vector. A CNN essentially consists of filters (convolutional layer) and aggregation layers

(pooling layer), which are alternately repeated, and at the end of one or more layers of "normal" completely connected neurons (dense / fully connected layer).

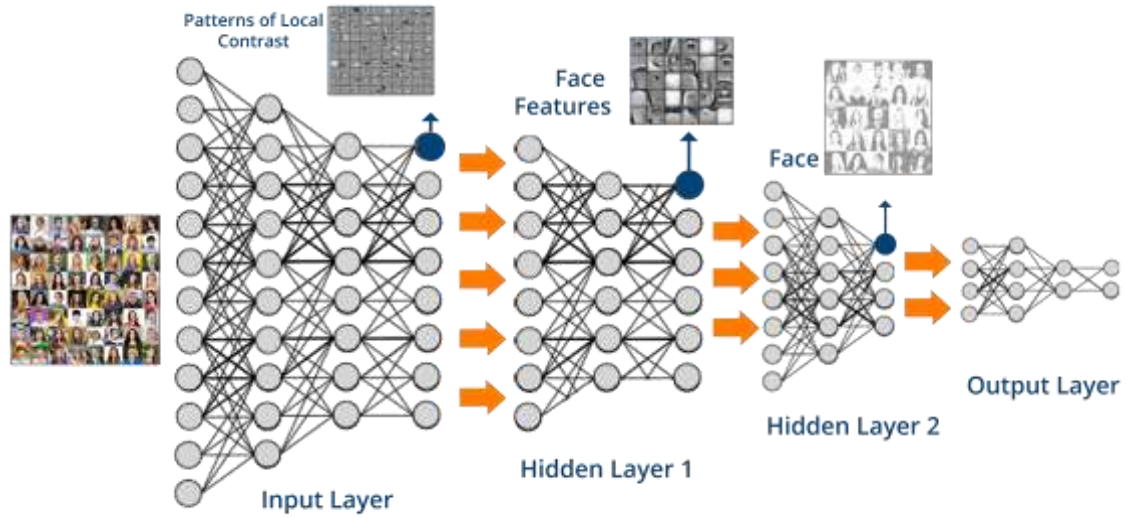


Figure 13: Convolutional Neural Networks

Figure 13 shows the convolutional neural networks applied for face recognition.

4.2.2 Support Vector Machine (SVM)

Support vector machines are methods for supervised machine learning, which are mainly used for binary classification. The training data are plotted in n -dimensional space and the algorithm tries to draw a limit with the greatest possible distance to the next sample [Heinert, 2010]. The limit is defined by means of the objects closest to it, which are therefore also called support vectors. Vectors that are far away from the limit are not decisive for the calculation. This is why they do not need to be loaded into main memory, which makes SV very memory efficient. A limit is a hyperplane. This is a subspace, the dimension of which is 1 smaller than its surroundings. For example, in three-dimensional space, a hyperplane would be a two-dimensional plane. And in two-dimensional space, a hyperplane would simply be a straight line.

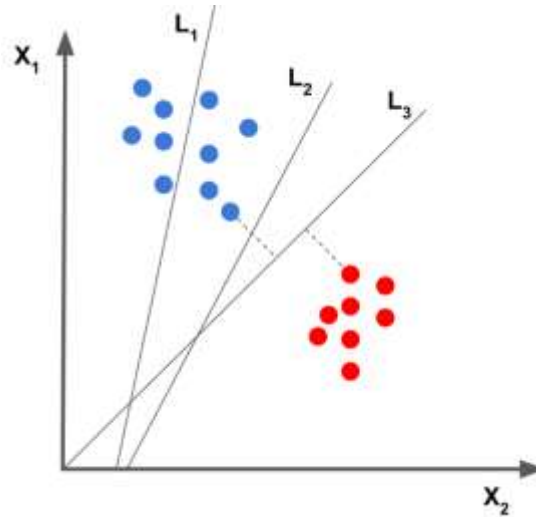


Figure 14: Classification by using SVM

4.2.3 Logistic Regression

Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables [Sperandei, 2014]. Logistic regression is a form of regression analysis that is used to predict a nominally scaled, categorical criterion. This means the logistic regression is used when the dependent variable has only a few, equal values. An example of a categorical criterion would be the outcome of an entrance examination in which one can only either be “accepted” or “rejected”. If the criterion in the logistic regression has only two values, then one speaks of a binary logistic regression. If, on the other hand, the criterion has more than two categories, the method is called multinomial logistic regression.

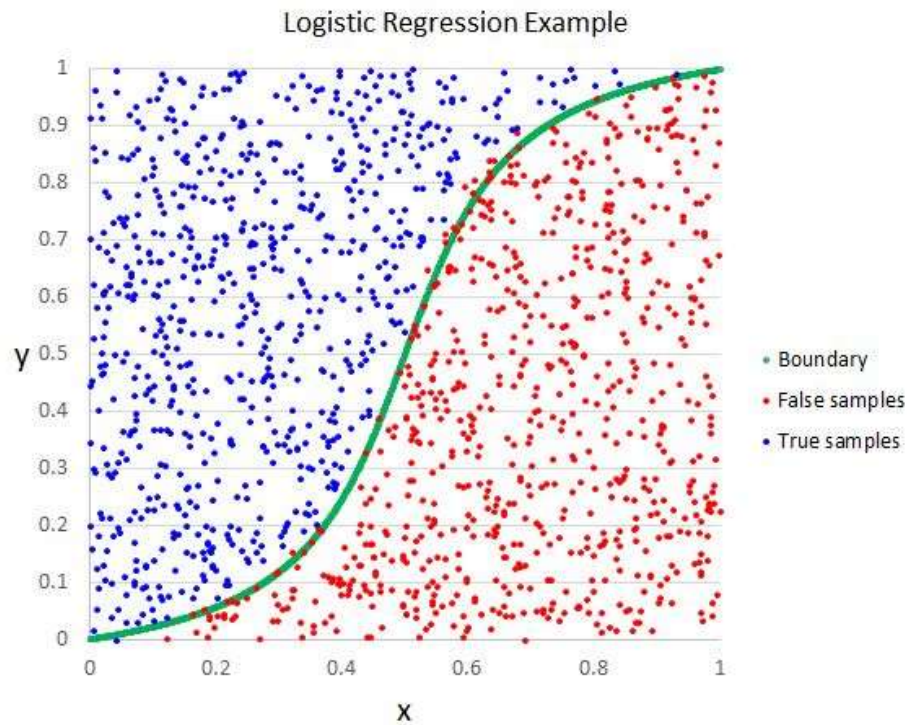


Figure 15: Logistic Regression

4.2.4 Naive Bayes

Naive Bayes is a machine learning technique by means of which objects, for example text documents, can be divided into two or more classes [Berrar, 2018]. The classifier learns by analyzing special training data for which the correct classes are given. The naive Bayesian classifier is used, among other things, when it comes to determining the probabilities of the classes on the basis of a series of observations. The model is based on the assumption that the variables are conditionally independent depending on the class.

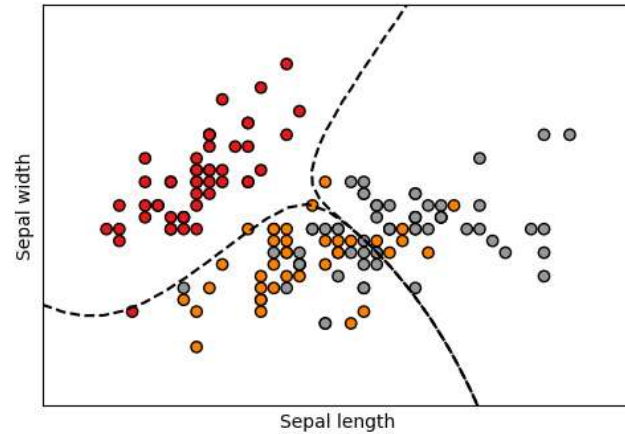


Figure 16: Naïv Bayes

4.2.5 Random Forest

A random forest is a classification algorithm consisting of many decision trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree [Cutler, 2011].

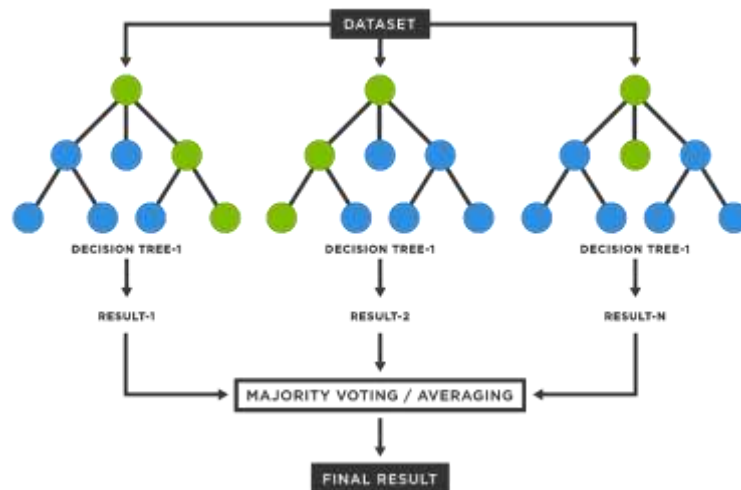


Figure 17: Random Forest

4.2.6 Decision stump

A decision stump is a machine learning model that consists of a single-level decision tree. That is, it is a decision tree with an internal node (the root) that is directly connected to the terminal nodes (its leaves) [Appaswamy, 2018]. A decision stump makes a prediction based on the value of only a single input feature. Sometimes they are also called 1 rule. Several variations are possible depending on the type of input attribute. For nominal features, one can build a stump that contains a leaf for each possible feature value, or a stump with two leaves, one of which corresponds to a selected category and the other of which corresponds to all other categories. For binary features, these two schemes are identical. A missing value can be treated as another category. Decision stumps are often used as components (“weak learners” or “basic learners”) in ensemble machine learning techniques such as bagging and boosting.

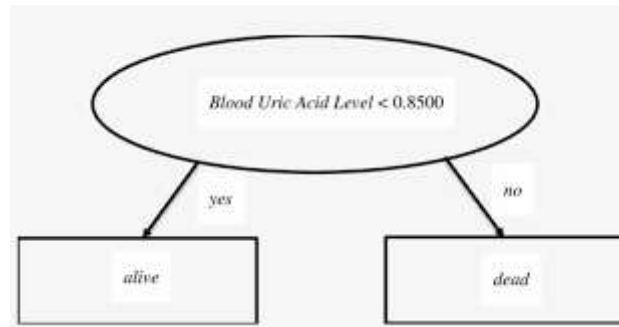


Figure 18: Decision Stump

4.2.7 AdaBoost

The algorithm trains several times and then the corresponding class is predicted by means of a majority vote of the algorithms [Chengsheng, 2017]. AdaBoost goes one step further and adapts the dataset in each iteration so that some incorrectly classified data points are classified correctly. According to Freund and Schapire (1999), AdaBoost does well on texts.

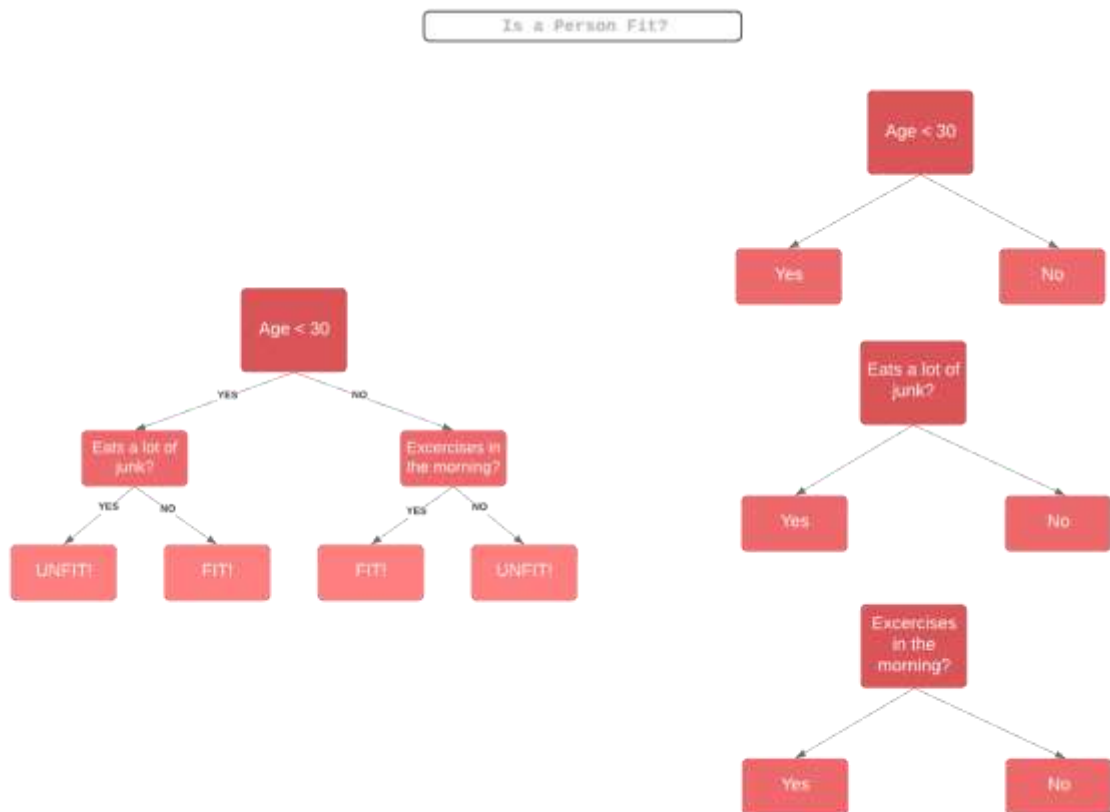


Figure 19: AdaBoost

4.2.8 Related application

Supervised machine learning can be used, for example, to predict house prices. A number of house attributes (also called features) such as living space, plot of land, number of rooms or schools in the vicinity must already be available for a number of houses. This data should influence the associated house prices either directly or indirectly. In addition, the house prices must be known for the existing data, which is also called training data, so that this relationship can be used for the creation (training) of a model. Once trained, the resulting model will be able to predict the house prices of new homes with some accuracy, as long as the information is available for the corresponding home features. With additional training data, a better adapted learning algorithm or a data cleaning step that removes incorrect or missing data, the prediction can be further improved automatically. For experienced data scientists, the last point is one of

the most important steps. In order to generate a good model, the data must be of the best possible quality.

Improved processes for image recognition make it possible for machine learning to evaluate image data more intensively. Visual information is extremely rich, is available very quickly and is usually available in very large quantities. Using more data can lead to better decisions.

Economical methods for image recognition are available:

- Object detection and tracking
- Location and position detection
- Size and shape measurement
- Surface inspection
- Pattern recognition
- Handwriting recognition
- Barcode recognition
- Completeness or fill level check
- Traffic control and guidance
- Person recognition

Chapter 5 PROPOSED NEW APPROACH

Figure 20 shows the proposed new approach, InnoCrowd. We analyze user requirement and product geometry sample as input dataset. Actually, these two aspects can be handled individually. For each of them we have developed an optimized knowledge database. In practice it is sufficient to give only text dataset or only image dataset. For the both use case the system will work as expected.

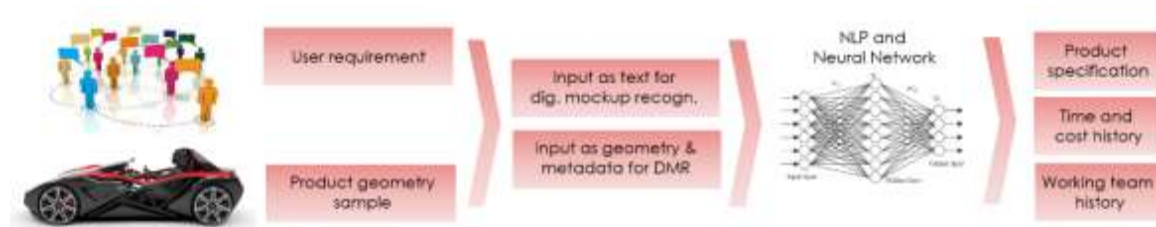


Figure 20: Approach of InnoCrowd

The first block of the figure above is representing the real live data from crowdsourcing platform. The crowdsourcer published their product requirement and product geometry sample in the crowdsourcing platform (website). The second block is representing the generation of dataset. The dataset was generated by using web scrapper as a data collection engine and windows script as data preprocessor. The result of this step is the dataset with the correct format that can be used for building up a knowledge database (the engine of InnoCrowd). The third block is representing the engine of InnoCrowd which consist of optimized algorithms for data classification by using NLP and machine learning. The last, fourth block is representing the output of InnoCrowd as a Decision Support System. It gives the related crowdsourced product development project from the past that match the new one. Product specification, time, cost and working team history are presented to the user for enabling allow risk decision. The next sub chapters explain the blocks in detail.

5.1 Generating the Dataset

A web scraper is used in order to collect datasets from the crowdsourcing platform. Web scraping, also called screen scraping, generally refers to the process of extracting, copying, saving and reusing external content in the network. In addition to manual scraping, in which content is copied by hand, some tools have also become established for the automated reading of websites. A positive use case of web scraping is the indexing of websites by Google or other search engines.

Various technologies and tools are used in web scraping:

- Manual scraping: In fact, both content and source code sections of websites are occasionally copied by hand. Internet criminals use this method especially when bots and other scraping programs are blocked by the robots.txt file.
- Software tools: Web scraping tools such as Scraper API. It enables the creation of web scrapers even with little or no programming knowledge. Developers also use these tools as a basis for developing their own scraping solutions.
- Text Pattern Matching: The automated comparison and reading of information from websites can also be done with the help of commands in programming languages such as Perl or Python.
- HTTP manipulation: Content from static or dynamic websites can be copied using HTTP requests.
- Data mining: Web scraping is also possible via data mining. For this purpose, web developers rely on an analysis of templates and scripts in which the content of a website is embedded. In doing so, they identify the content they are looking for and play it on their own page using a so-called wrapper.
- HTML parser: The HTML parsers known from browsers are used in web scraping to read out and convert the content searched for.

The process of web scraping is straightforward, although implementation can be complex. First, the piece of code that is used to retrieve the information, which is a scraper bot, sends an HTTP GET request to a specific website. When the website responds, the scraper analyzes the HTML document for a certain data pattern. Once the data is extracted, it is converted into any specific format designed by the scraper bot's author. Scraper bots can be designed for many purposes, e.g.: Content scraping - content can be extracted from the website to replicate

the unique benefit of a particular product or service that relies on content. For example, a product like Yelp relies on reviews. A competitor could scrape all of Yelp's review content, reproduce the content on their own website, and pretend that the content is original. Price scraping - By scraping price data, competitors can gather information about their competition. This allows them to gain a unique advantage. Contact scraping - Many websites contain email addresses and phone numbers in plain text. By scraping pages like an online employee directory, a scraper can collect contact information for bulk mailing lists, robot calls, or malicious social engineering attempts. This is one of the main ways that both spammers and scammers discover new targets.

From the given available tools above we choose the second option, namely a web scraping tool. This is because it can cover our requirement easily. It can collect text and image files automatically after a quick setup of the tool.

We select areas on the website which information are relevant to be collected. For example, on the area where the product requirement is shown on the website. And, only on the area where images are located or linked on the website.

5.1.1 Collecting the dataset

There are many crowdsourcing providers with different topic coverages and different task types. The following are the most famous providers.

Provider	Task Type	Coverage
AMAZON MECHANICAL TURK	micro work	all topics
TOPCODER	macro work	software & engineering
UPWORK	macro work	all topics
CLICKWORKER	micro work	all topics
ENNOMOTIVE	macro work	engineering
JOVOTO	macro work	all topics
EMBRIOO	macro work	all topics
INNOCENTIVE	complex project	engineering
GRABCAD	macro work	engineering
LOCAL MOTORS	complex project	engineering
CADCROWD	macro work	engineering

Crowdsourcing provider which provides micro work and for our research we collect the dataset from two crowdsourcing provider, namely CADCROWD and TOPCODER.

CADCROWD is an online matching service that helps companies hire global CAD (Computer-Aided Design) workers, literally saving tens of thousands of dollars on their CAD drafting and engineering work. CADCROWD allows companies to hire contract workers on demand by posting their contract job on cadcrowd.com, receiving bids, and managing the contract through CADCROWD. It has over 50,000 quality certified global staff that can be hired on-demand. It enables clients to achieve cost savings of 70-85% on CAD work and is an important solution to the upcoming labor shortage for CAD workers.

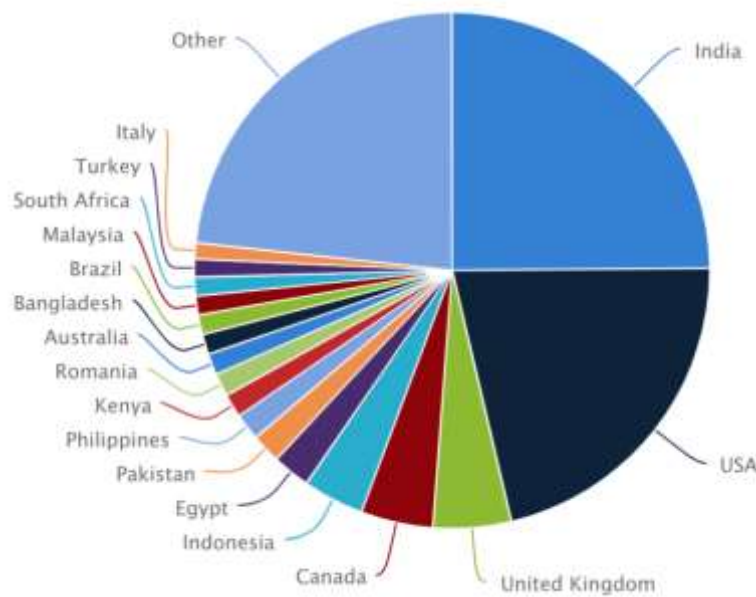


Figure 21: Global CADCROWD designers and engineers

A large part of the work has been focused around helping clients create designs and 3D models for new products ideas. Clients take the work created by freelancers to manufacturers and/or use the designs for crowd funding campaigns on sites like Kickstarter.

Long term goal of CADCROWD is to build very large reliable network of freelancers for 3D modeling and CAD design work. A large number of the work on the site has been completed through contests (crowd sourcing), where clients pick the design that best suits their idea. The work showcase and reputation (via points system leaderboard) has allowed top designers and engineers to earn a significant amount of money through private projects and ongoing assignments with clients. So far, a large number of design contests being posted and their

number of designers has grown steadily. Whereas some businesses focus on building a community or library of CAD models, it appears that CADCROWD is focusing on building a library of CAD designers. And unlike other freelance job boards, it has found a way to provide a positive experience for both the client and customer while allowing several methods of design completion: through competitions, by hourly labor, or within private projects.

TOPCODER is a crowdsourcing company with an open global community of designers, developers, data scientists, and competitive programmers. TOPCODER pays community members for their work on the projects and sells community services to corporate customers, medium-sized and small businesses. TOPCODER also organizes the annual TOPCODER Open tournament and a number of smaller regional events.

The TOPCODER community is the main source of manpower behind all TOPCODER projects. It is open and global: anyone can join and compete with some legal restrictions dictated by US laws and listed in the community terms, with no financial obligations on TOPCODER. Participation in challenges that are organized in the interest of commercial customers also usually requires the community member to sign a non-disclosure agreement. Intellectual property for the winning entries in commercial competitions will be given to the customer in exchange for cash prizes paid to the winners. While the majority of community members participate in TOPCODER challenges as regular competitors, those recognized for their performance and involvement in community life (through communicating in TOPCODER forums, attending TOPCODER events, etc.) will receive additional special roles offered in the community. These include: copilots (technical coordinators of challenges), problem writers, reviewers, etc. From late 2014 to late 2017, a Community Advisory Board (CAB) was selected from active community members for a year to term communication between the top coder -Improve the company and its community. In 2018, the CAB was replaced by the TOPCODER MVP (Most Valuable Player) program. There are four main segments of every TOPCODER community that are open to every member: Design, Development, Data Science and Competitive Programming. In addition, since the end of 2017, it has been creating sub-communities as part of its hybrid crowd offering that are dedicated to specific customers / projects. The sub-communities can require members to meet additional eligibility criteria before joining.

The dataset for the experiment was extracted from the crowdsourcing platform CADCROWD and TOPCODER. More than 900 crowdsourcing projects are listed in the platform CADCROWD. More than 800 projects were finished and therefore could be used as input for our knowledge database in InnoCrowd. The dataset consists of text and images. We extracted the data with a web scraper. At the beginning we processed the text and image data separately in the data mining tool “Orange.”.

We collected all necessary information from the crowdsourcing platform:

- As text, the following information:
 - Extraction of product requirement defined by the crowdsourcer
 - Information on rewarding
 - Information on project timeline
 - Information on crowd, who were the winner and the solution provider
- As image, the following information
 - Image files given by the crowdsourcer
 - Image files submitted by the crowd as contribution

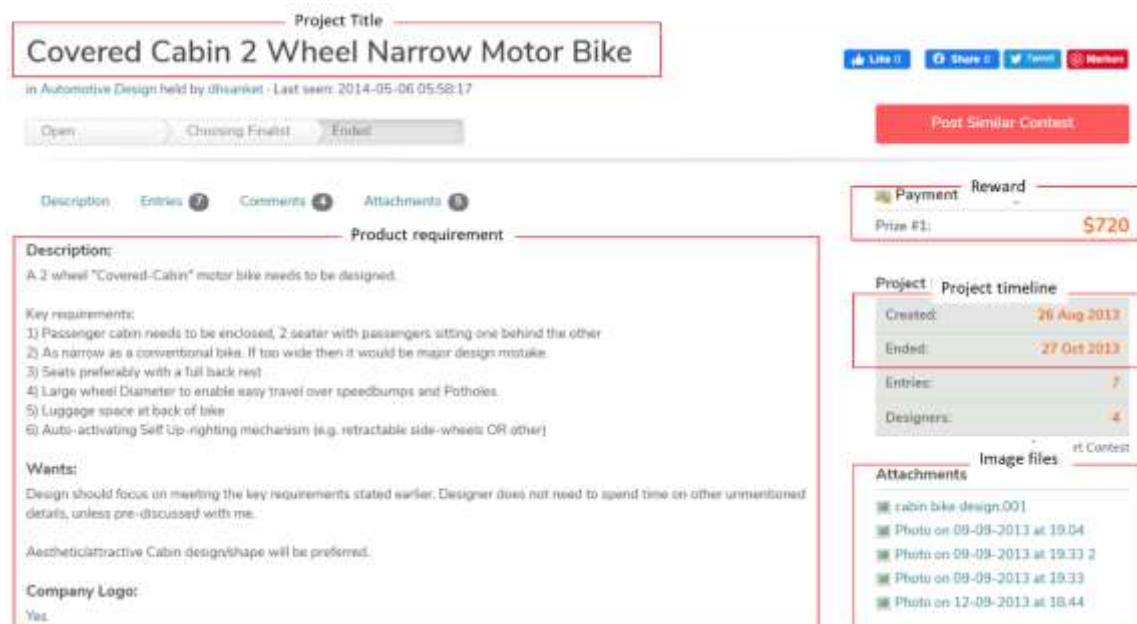


Figure 22: Dataset from a crowdsourcing platform (Source: CAD Crowd)

Figure 22 shows a sample of data to be collected from a crowdsourcing platform. If we open image files from the link give on that site, we will see the following images (in Figure 23) as imagination given by the crowdsourcer on how the product may look like.



Figure 23: Images given by the crowdsourcer (Source: CAD Crowd)

If we scroll down on the same page of the website, then we will find contributions from the crowd. Figure 24 shows contributions from 4 member of the crowd. Here we collect the images and information about the owner of the image.

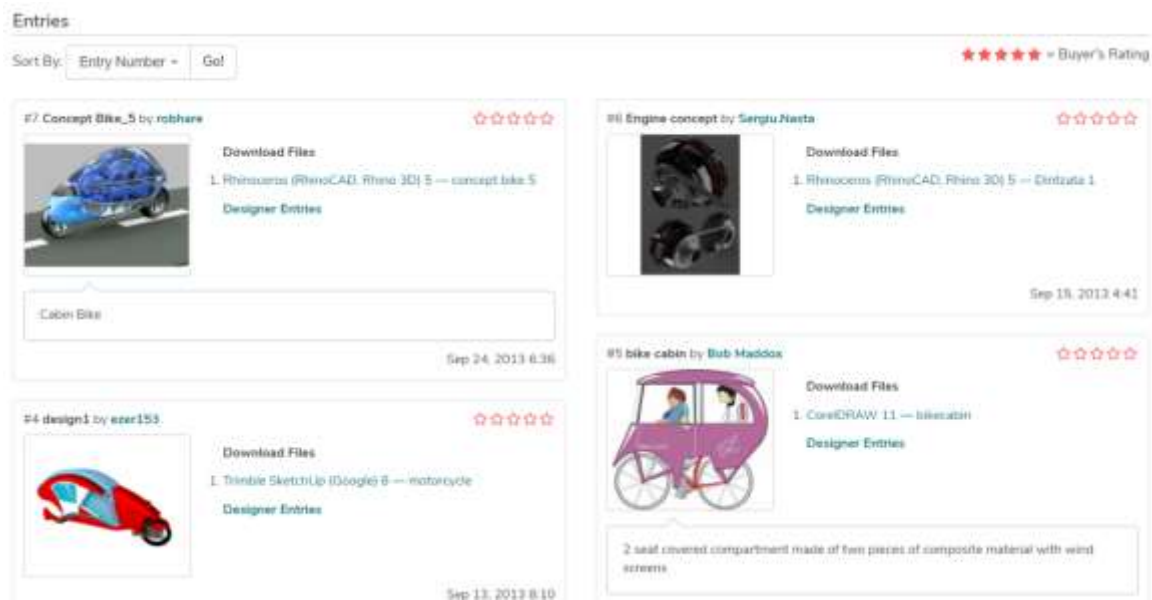


Figure 24: Contributions from the crowd (Source: CAD Crowd)

Depends on the response from the crowd, each crowdsourcing project may get not only four contributions but it could reach more than 100 contributions.

5.1.2 Preprocessing of the dataset

It took several hours to get the result from the web scrapping process. The rough dataset is then available in the format of Microsoft Excel Table and a folder fulfilled with image files. The data has to be preprocessed before being used in the classification system. The texts are grouped into several segments. We have found out that some segments are not relevant for the classification task. Therefore, we cut this from the dataset. We give specific group naming to some segment so that it gives more semantic meaning which is relevant for the classification task. Among others the group name “Dimension”, “IT specification and “General specification”. The table consist of two tabs. One tab is the collection of text from the selected text area on the website. The other tab is the collection of image links which relate information of the crowdsourcing project with the name of image files collected in the image folder.

Figure 25 shows the first tab of the table. We can see there that big spaces are existing between text. This is because the web crawler copied also spacing between paragraphs. Off course this space is not relevant for the next step during the text analysis. Therefore, we have to preprocess this data from excel table by using text editor Notepad ++. As a result is a text dataset with the conforming format for the step of text analysis with NLP.

	A	B	C	D	E	F	G	H	I	J	K	L	M					
1	web-scraped-order	web-scraped-start-url	page	page-href	title	title-href	desc	want	Dontwant	Sample	Prize	created	ended	entries	designers	Winner	category	image1-src
2	1609884273-12639	"https://www.cadcrowd.com/projects/status-finished&page=7"	"Hi Cad Crowders!"															"Help us design our sustainable barbecue!" "https://www.cadcrowd.com/contest/235"
3																		
4																		
5																		
6																		
7																		We would like to ask you for your help to bring our sustainable barbecue concept to a reality. We believe that we can develop a barbecue that has a lower environmental footp
8																		
9																		
10																		
11																		Our design is based on a simple principle, that is widely used in African villages: the Rocket Stove (see: www.aidafrika.net/rocket_stoves/). Rocket Stoves use a third of the fue
12																		
13																		
14																		
15																		The rocket stove is not so common in the western world, as they often look cheap, simple and don't have the same aura as a Green Egg or Olyr. We want to change that and dev
16																		
17																		
18																		
19																		Currently, we already have a vision of what the barbecue should look like, but we need to bring our idea's to 3D models that we can use to 1) promote our product among inv

Figure 25: The collected text from the crowdsourcing platform

The most time-consuming process was the process for collecting the images. Although we collect images with mainly small sizes, but the amount of images to be collected was very large. We can imagine that each crowdsourcing project may related with more than 50 image files. All together will give a really great number. Figure 26 shows the sample of the collected image in image folder.

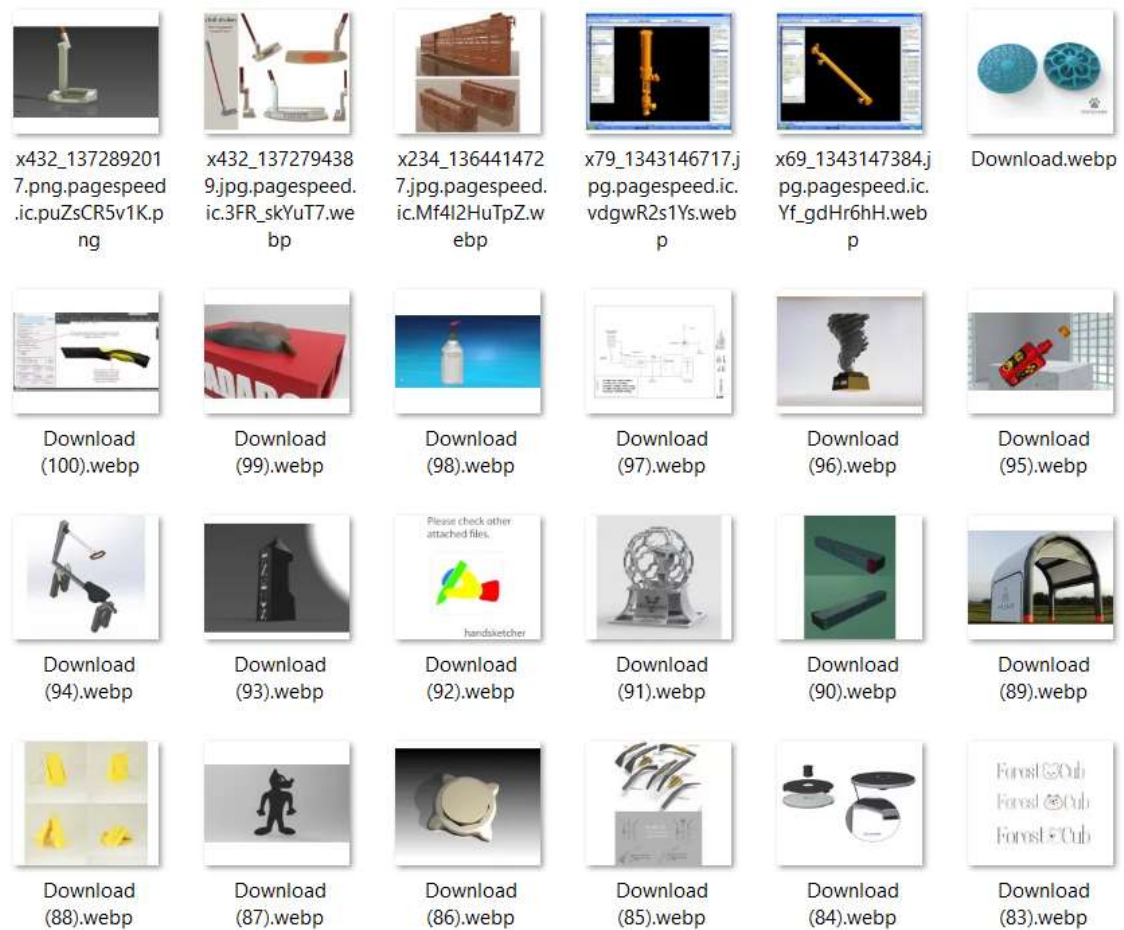


Figure 26: Collected image files

Also, during the collection of the image files, we needed to do some preprocessing works before we can use it for the next step during the image analysis. The image link given in the second tab of the table created by the web scrapper is to be matched with the text information from the first tab of the table. There is an ID number which relates the both information. We

took this ID number to relate the project with the image name on the image folder. We created a windows script for executing this task.

5.2 Concept for Text Analysis and Classification

On the chapter 4 we provide state of the art among others for the text analysis, namely by using NLP. We are using NLP and optimize it for the best analysis and classification performance related to our dataset. Optimizing here means:

- Selecting the appropriate tool
- Selecting the best algorithm
- Upgrade the performance by adjusting the parameter during the experiment
- Re-preprocess the dataset and re-adjust the parameter when necessary

5.2.1 Selecting the appropriate tool

For speech processing with NLP, there are a few tools and applications that can help. The most used tools are among others Python and Natural Language Toolkit (NLTK), Statistical NLP and Orange - Data Mining.

Python and Natural Language Toolkit (NLTK). The well-known programming language Python offers tools for dealing with NLP-specific tasks. Some of these can be found in the Natural Language Toolkit (NLTK). This is an open-source collection of programs, libraries and other resources for creating NLP programs. The Natural Language Processing Toolkit contains some tools for handling the functions and areas of application. There are also libraries for subtasks such as word segmentation, sentence parsing or stemming and lemmatization to understand a text.

Statistical NLP. The first NLP applications included hand-coded, rule-based systems. Although these could fulfill certain tasks, they were not useful for a wide range of tasks. Statistical NLP combines computer algorithms with machine learning as well as deep learning models to automatically extract and classify text and language data and then to assign a possible meaning to the elements with a statistical probability. In the meantime, deep learning models

offer the possibility of creating NLP systems that learn independently during the process. This allows these systems to develop increasingly precise solutions over time.

Orange – Data Mining is an open-source data visualization, machine learning and data mining toolkit. It features a visual programming front-end for explorative rapid qualitative data analysis and interactive data visualization. Features of orange for NLP are among others: text preprocessing, text clustering and text classification. The user can use predictive models to classify documents by authorship, their type, sentiment and so on. In this workflow Orange classifies documents by their Aarne-Thompson-Uther index, that is the defining topic of the tale. It uses two simple learners, Logistic Regression and Naive Bayes, both of which can be inspected in the Nomogram.

The Statical NLP has some really good features such as deep learning for NLP. It can give a high quality of text analysis. But when comparing Statical NLP with NLTK then it has lower performance than NLTK. NLTK is the most used tool for NLP in research area, because of the comprehensive library it has. Very large community is developing NLTK further.

We compared NLTK with Orange-Data mining tool. When considering the available features and library then NLTK is the winner because it covers basic and the most updated topics of NLP. When considering the usage of the tool then Orange is the winner. It is really easy to setup the data processing analysis chain when using Orange. We can do drag and drop of the existing library. The feature for data preprocessing is also available and very useful. Roughly we can say, the development and experiment of NLP by using Orange can be roughly two time faster than using NLTK. It is because NLTK is to be implemented per command line, which is not always user friendly. On the other side, when some features are missing in Orange, then we can implement a NLTK script and add this to the process chain in Orange. That means, by using Orange we can use also NLTK at the same time. Therefore, we use Orange for text analysis in this PhD work.

Orange is available for Windows, Mac OS X and various Linux systems. Databases or simple data structures can be visualized and evaluated with just a few clicks. Due to its simple structure, Orange is particularly interesting for newcomers. But it also offers many exciting features for advanced users. Easily design complex evaluations Using widgets, basic functions such as the reading of data, the display of tables or the visualization of data can be carried out. In addition, individual widgets can be developed in Orange that are tailored to the respective

problem of the user. Operation is user friendly thanks to drag & drop. Complex tools can be easily dragged and linked into the Orange canvas. In addition, changes are directly traceable thanks to the interactive visualization. After a short time to get used to the symbols and the various combination options of the tools, the operation is very intuitive. A small point of criticism is that the orange canvas becomes confusing with complex queries with several visualizations. However, due to the generally simple handling, one can overlook this error. There is a large selection of options for visualization, ranging from bar charts and line charts to scatter charts. Orange is also suitable for people with existing programming skills. The program has the option of running Python scripts via a terminal window and offers programmers development environments such as PyCharm or Shells such as iPython.

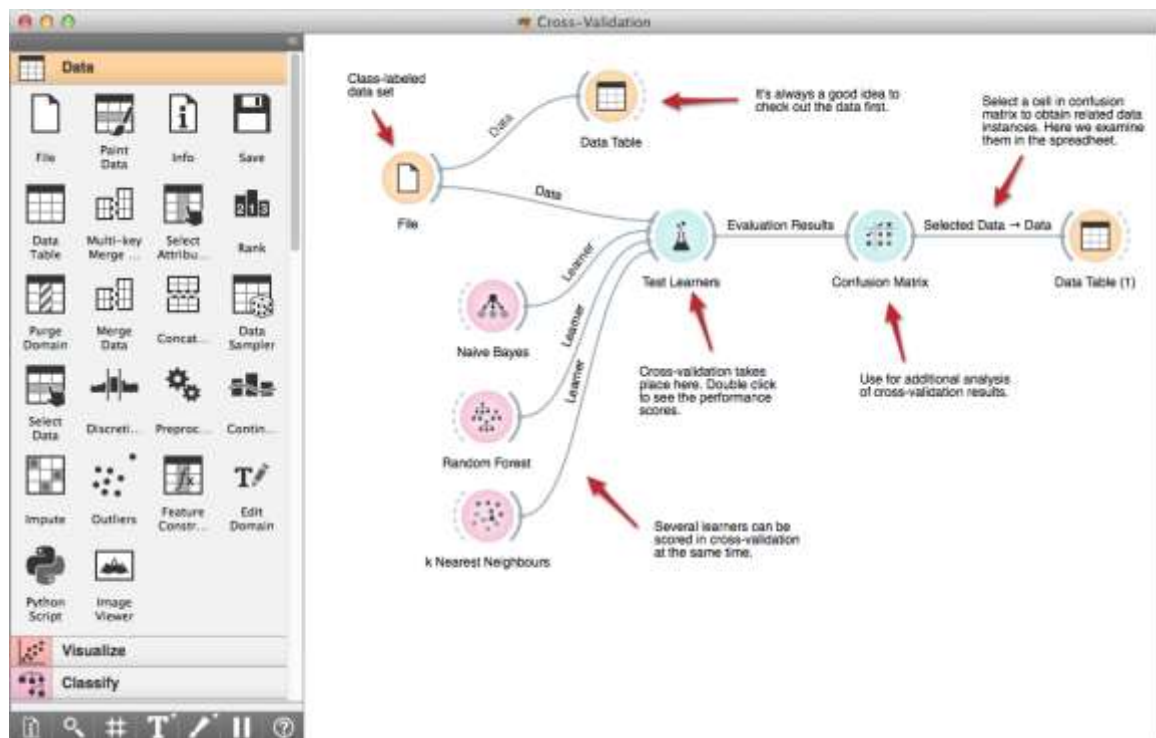


Figure 27: Orange – Data Mining

5.2.2 Process chain for text analysis

We developed a process chain for the text analysis. Figure 28 shows the process chain. It starts with the collected text dataset which is called as a corpus. And then the text is preprocessed in order to eliminate unnecessary text components. The is transformed into numbers by using

the feature “Bag of words” [Zhang, 2010]. We used 6 machine learning algorithms and compared their performance for the text classification task. The result of the comparison can be seen on the feature “Test and Score”.

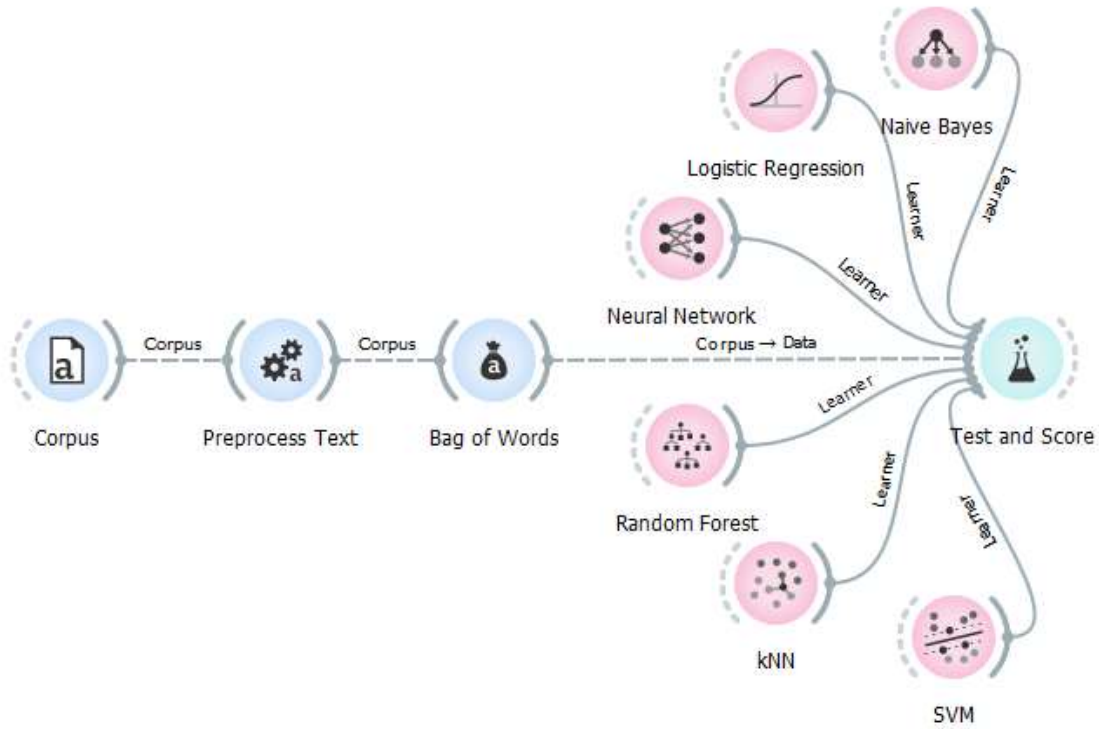


Figure 28: Process chain for text analysis

5.2.2.1 Corpus

The process is initiated in the first step called “Corpus” which takes the text file in a specific data format. Corpus is a collection of text which can consist of several documents. In case of our dataset each crowdsourcing project is a single document.

Corpus widget can work in two modes:

- When no data on input, it reads text corpora from files and sends a corpus instance to its output channel. History of the most recently opened files is maintained in the widget. The widget also includes a directory with sample corpora that come pre-installed with the add-on. The widget reads data from Excel (.xlsx), comma-separated (.csv) and native tab-delimited (.tab) files.

- When the user provides data to the input, it transforms data into the corpus. Users can select which features are used as text features.

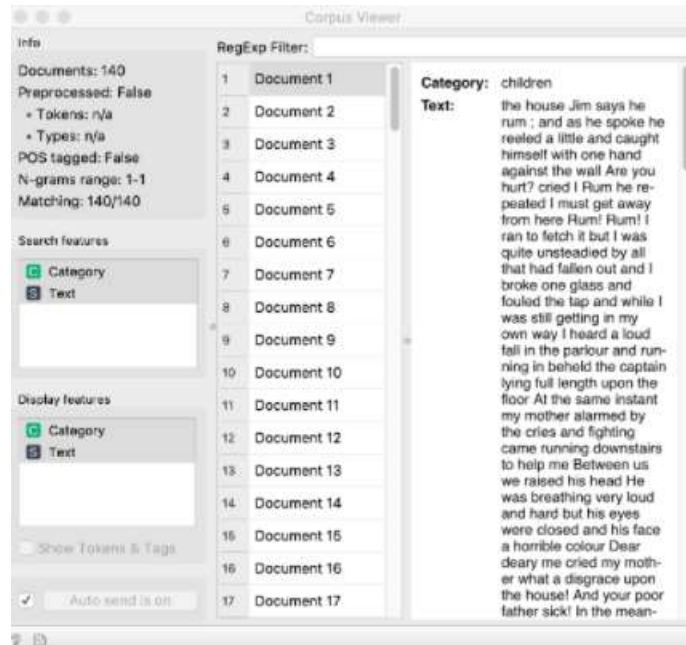


Figure 29: Each crowdsourcing project is represented as a single document in Corpus

Before inserting the text into corpus, we had to make sure that the text data format is in compliance with the widget Corpus. As explained in chapter 5.1.2 we had at the beginning the text data in excel table. We preprocess the text by using Notepad++ to eliminate unnecessary spaces and save the text file in txt format. We had also to place divisions of the text in the right position. We implement a script for executing this task. However, this preprocessing step is before the analysis in Orange. Therefore, the next sub chapter deals with preprocessing of text inside the Orange, which is a different step.

5.2.2.2 Preprocess Text

It is necessary to convert the text file extracted from the web into the correct data format. The second step “Preprocess Text” is the step for handling parts of text that are not relevant for the data mining. For an example, commas, dots, low level words such as “and”, etc. Preprocess Text splits the text into smaller units (tokens), filters them, runs normalization (stemming,

lemmatization), creates n-grams and tags tokens with part-of-speech labels. Steps in the analysis are applied sequentially and can be reordered. The preprocessing step includes: Transformation, tokenization, normalization and filtering.

The screenshot displays the NLTK preprocessing interface, which is organized into four main sections, each with a numbered icon in the top right corner:

- Transformation (2):** Contains four checkboxes: ☒ Lowercase, ☐ Remove accents, ☐ Parse html, and ☐ Remove urls.
- Tokenization (3):** Contains five radio buttons: Word_Punctuation, Whitespace, Sentence, Regexp (selected), and Tweet. A text field labeled "Pattern:" contains the value `\w+`.
- Normalization (4):** Contains four radio buttons: Porter Stemmer (selected), Snowball Stemmer, WordNet Lemmatizer, and UDPipe Lemmatizer. The Snowball Stemmer and UDPipe Lemmatizer options have a "Language:" dropdown menu set to "English". The UDPipe Lemmatizer option also includes an unchecked checkbox for "UDPipe tokenizer".
- Filtering (5):** Contains several options:
 - ☒ Stopwords: Includes a language dropdown (English), a filter type dropdown (none), and buttons for file selection and refresh.
 - ☐ Lexicon: Includes a filter type dropdown (none) and buttons for file selection and refresh.
 - ☐ Regexp: Includes a text field with a complex regular expression pattern.
 - ☐ Document frequency: Includes two radio buttons, "Relative" (selected) and "Absolute". The "Relative" option has two range sliders set to 0.10 and 0.90. The "Absolute" option has two range sliders set to 1 and 10.
 - ☐ Most frequent tokens: Includes a range slider set to 100.

Figure 30: Preprocessing text

Transformation transforms input data. It applies lowercase transformation by default. Lowercase will turn all text to lowercase. Remove accents will remove all diacritics/accents in text. naïve → naïve. Parse html will detect html tags and parse out text only. <a href...>Some

text → Some text. Remove urls will remove urls from text. This is a <http://orange.biolab.si/> url. → This is a url.

Tokenization is the method of breaking the text into smaller components (words, sentences, bigrams). Word & Punctuation will split the text by words and keep punctuation symbols. This example. → (This), (example), (.). Whitespace will split the text by whitespace only. This example. → (This), (example.). Sentence will split the text by full stop, retaining only full sentences. This example. Another example. → (This example.), (Another example.). Regexp will split the text by provided regex. It splits by words only by default (omits punctuation).

Normalization applies stemming and lemmatization to words. (I've always loved cats. → I have always love cat.) For languages other than English use Snowball Stemmer (offers languages available in its NLTK implementation) or UDPipe. Porter Stemmer applies the original Porter stemmer. Snowball Stemmer applies an improved version of Porter stemmer (Porter2). Set the language for normalization, default is English. WordNet Lemmatizer applies a networks of cognitive synonyms to tokens based on a large lexical database of English. UDPipe applies a pre-trained model for normalizing data.

Filtering removes or keeps a selection of words. Stopwords removes stopwords from text (e.g. removes 'and', 'or', 'in'...). Select the language to filter by English is set as default. You can also load your own list of stopwords provided in a simple *.txt file with one stopword per line. Click 'browse' icon to select the file containing stopwords. If the file was properly loaded, its name will be displayed next to pre-loaded stopwords. Change 'English' to 'None' if you wish to filter out only the provided stopwords. Click 'reload' icon to reload the list of stopwords. Lexicon keeps only words provided in the file. Load a *.txt file with one word per line to use as lexicon. Click 'reload' icon to reload the lexicon. Regexp removes words that match the regular expression. Default is set to remove punctuation. Document frequency keeps tokens that appear in not less than and not more than the specified number / percentage of documents. Absolute keeps only tokens that appear in the specified number of documents. E.g. DF = (3, 5) keeps only tokens that appear in 3 or more and 5 or less documents. Relative keeps only tokens that appear in the specified percentage of documents. E.g. DF = (0.3, 0.5) keeps only tokens that appear in 30% to 50% of documents. Most frequent

tokens keeps only the specified number of most frequent tokens. Default is a 100 most frequent tokens.

5.2.2.3 Bag of Words

The third step “Bag of Words” is the step where the words in each dataset are counted and then converted into numbers in order to make it calculatable and comparable during the data classification in the next step. We consider a text as a set (a bag) of words [Zhang, 2010]. So a bag of words.

We take an example to understand this concept in depth. "It was the best time" "It was the worst time" "It was the age of wisdom" "It was the age of stupidity".

We'll treat each sentence as a separate document and make a list of all the words from all four documents except for punctuation. We get, "It", "was", "the", "best", "of", "times", "worst", "age", "wisdom", "stupidity".

The next step is the creation vectors. Vectors convert text that can be used by the machine learning algorithm. We take the first document - "It was the best time" and check the frequency of the words out of the 10 unique words.

"It" = 1 "was" = 1 "that" = 1 "the best" = 1 "from" = 1 "times" = 1 "the worst" = 0 "Age" = 0 "Wisdom" = 0 "Stupidity" = 0

The rest of the documents read:

"It was the best time" = [1, 1, 1, 1, 1, 1, 0, 0, 0, 0]

"It was the worst time" = [1, 1, 1, 0, 1, 1, 1, 0, 0, 0]

"It was the age of wisdom" = [1, 1, 1, 0, 1, 0, 0, 1, 1, 0]

"It was the age of stupidity" = [1, 1, 1, 0, 1, 0, 0, 1, 0, 1]

In this approach, each word or token is referred to as a "gram". Creating a vocabulary from two pairs of words is called the bigram model. For example, in the first document, the bigrams read "It was the best time" as follows: "It was" "was the" "the best" "the best" "currently".

hidden	category	text	(...)
1	children	the house Jim s...	broke=1.000, by=4.000, trebly=1.000, basin=3.000, executed=1.000, picture=1.000, se...
2	children	has lived rough...	golden=1.000, carried=1.000, bar=2.000, confessions=1.000, air=1.000, again=5.000, t...
3	children	Now boy he sai...	gathering=1.000, letter=1.000, bring=1.000, resolved=1.000, payment=1.000, peculiar...
4	children	thanks to you b...	despair=1.000, thanks=1.000, finely=1.000, swift=1.000, terrors=1.000, rogues=1.000, ...
5	children	the empty ches...	curiosity=1.000, drag=1.000, retreat=1.000, beyond=1.000, brief=1.000, cowardice=1.000, ...
6	children	stood irresolute...	dance=3.000, furious=1.000, such=1.000, matter=1.000, fools=1.000, nearest=1.000, p...
7	children	WE rode hard al...	son=1.000, rascal=1.000, smoke=1.000, proud=1.000, hearty=1.000, villains=1.000, co...
8	children	same as the tatt...	entry=1.000, roll=1.000, cache=1.000, blank=1.000, rank=1.000, manned=1.000, houn...
9	children	IT was longer t...	transparent=1.000, housekeeper=1.000, explored=1.000, fancies=1.000, plans=1.000, ...
10	children	treasure Long J...	dream=1.000, picked=1.000, telescope=1.000, substance=1.000, unearthed=1.000, ro...
11	children	We are so grate...	whatever=1.000, favor=1.000, therefore=1.000, beam=1.000, dismay=1.000, dwelt=1.000, ...
12	children	I am told said t...	loudly=1.000, frock=1.000, bread=2.000, brook=1.000, around=1.000, grieve=1.000, g...
13	children	to find the one ...	watched=2.000, chin=1.000, menily=1.000, earnestly=1.000, stalks=1.000, stop=1.000...
14	children	take away the o...	unfriendly=1.000, nest=1.000, bites=1.000, truliv=1.000, oarty=1.000, lonesome=1.000...

Figure 31: Sample of the generated Bag of Words

When it comes to the thematic classification of texts, the Bag of Words model works very well. We do not count all words, only the words that have a thematic reference. Stop words occur frequently in all texts, regardless of the topic. Other function words do not contribute much to the correct classification either. Function words are the words that belong to a closed class of words. Closed word classes are those classes that consist of a relatively small fixed set of words, such as article, preposition or auxiliary verb. Bag of Words model creates a corpus with word counts for each data instance (document). The count can be either absolute, binary (contains or does not contain) or sublinear (logarithm of the term frequency).

5.2.2.4 Machine learning algorithms for text analysis

The fourth step is the testing and scoring process by using several algorithms. As seen in the picture, six algorithms are evaluated in order to get the best classification result. These algorithms are: Neural Network, Logistic Regression, Naïve Bayes, Random Forest, k-nearest neighbors (kNN) and Support Vector Machine (SVM).

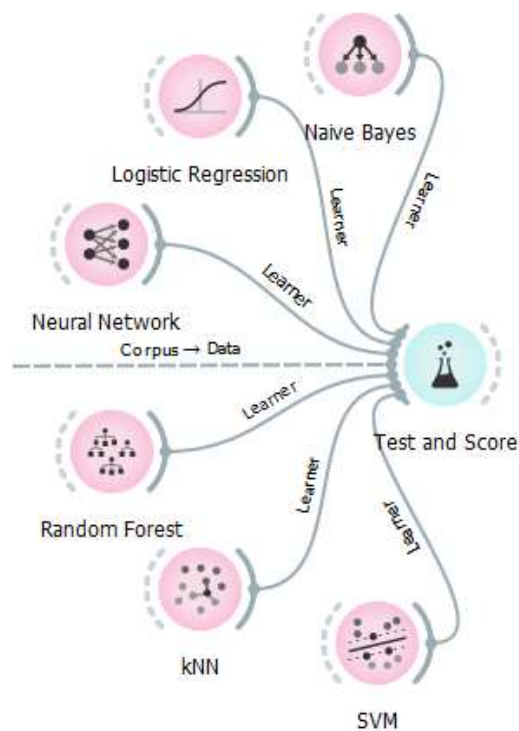


Figure 32: Machine learning algorithms for text analysis

At the beginning we used the given default parameter for each algorithm. We wanted to have a simple review of the performance of each algorithm. The default parameters for each algorithm are the following:

- Neural Networks (CNN)
 - Neurons per hidden layer: 100
 - Activation: ReLu
 - Solver: Adam
 - Alpha: 0,00010
 - Max iterations: 200
- Support Vector Machine (SVM)
 - Cost: 1,00
 - Regression loss epsilon: 0,10
 - Regression cost: 1,00
 - Complexity bound: 0,50

- Kernel: RBF
 - Numerical tolerance: 0,001
 - Iteration limit: 100
- Logistic Regression
 - Regularization type: Ridge (L2)
 - C strength= 1
- Naive Bayes
- Random Forest
 - Number of trees: 10
 - Growth control: do not split subset smaller than 5
- k-Nearest Neighbours (kNN)
 - Number of Neighbours: 5
 - Metric: Euclidean
 - Weight: Uniform

In order to give a better understanding on parameter of neural network that we want to optimize we give detail explanation of each parameter in the following paragraphs.

Number of Layer of Neural Network

In simplified terms, the structure of an NN can be imagined as follows: The model of the neural network consists of nodes, also called neurons, which receive information from other neurons or from outside, modify it and output it as a result. This is done using three different layers, each of which can be assigned a type of neuron: those for the input (input layer), for the output (output layer) and the so-called hidden neurons (hidden layers).

The information is received by the input neurons and output by the output neurons. The hidden neurons lie in between and depict internal information patterns. The neurons are connected to one another via so-called edges. The stronger the connection, the greater the influence on the other neuron.

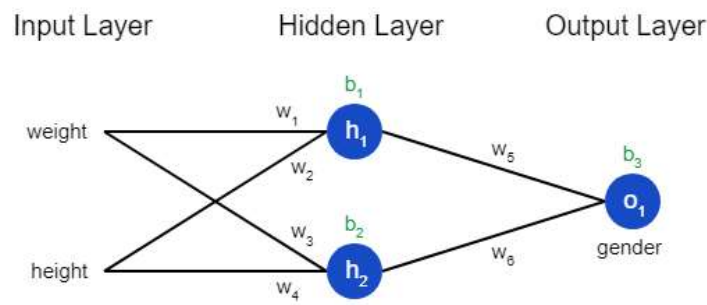


Figure 33: Layers of neural network

Three kinds of layer:

- **Input layer:** The input layer supplies the neural network with the necessary information. The input neurons process the data entered and pass them on to the next layer in a weighted manner.
- **Hidden Layer:** The hidden layer is between the input layer and the output layer. While the input and output layers only consist of one level, there can be any number of levels of neurons in the hidden layer. Here the received information is weighted again and passed on from neuron to neuron to the output layer. The weighting takes place at each level of the hidden layer. However, the exact processing of the information is not visible. Hence the name, hidden layer. While the incoming and outgoing data are visible in the input and output layer, the inner area of the neural network is basically a black box.
- **Output layer:** The output layer is the last layer and immediately follows the last layer of the hidden layer. The output neurons contain the resulting decision, which emerges as the flow of information.

Activation

The activity function (transfer function, activation function) represents the relationship between the network input and the activity level of a neuron. The activity function is visualized in a 2-dimensional diagram, with the network input of the unit on the abscissa (x-axis) and the ordinate (y Axis) the corresponding activity level is displayed. The activity level is then

transformed into the output by a so-called output function, which the neuron sends on to other neurons. The identity function is often used as the output function, i.e. the output is equal to the activity level.

A distinction is made between different activity functions:

- Linear activity function: Here the relationship between network input and activity level is linear.
- Linear activity function with threshold: Before the relationship between the two quantities becomes linear, a previously defined threshold must be exceeded. This can be useful if a network input that is too low (e.g. noise) should not be passed on as a signal.
- Binary threshold function: There are only two states of the activity level, 0 (or sometimes -1) or 1.
- Sigmoid activity function: This type of activity function is used in most models that simulate cognitive processes. A distinction can be made between the logistic function and the hyperbolic tangent function. However, both functions behave relatively similarly: If the network input (in terms of amount) is large and negative, then the activity level is close to 0 (logistic function) or -1 (tangent hyperbolic function), then initially increases slowly (a kind of threshold), after which the increase becomes steeper and resembles a linear function. With a high network input, the value then asymptotically approaches 1.

Solver

A very common method in the field of neural networks is the calculation of the adaptive learning rate adjustment using the “Adaptive Moment” Estimation” function (Adam)

Further methods for gradient-based optimization of the learning rate adjustment are: Nesterov accelerated gradient, Adagrad, Adadelata, RMSprop, AdaMax, Nadam and AMSGrad

Regularization

For the development of a network, the error is derived from the data extended by a term that takes the complexity of the model into account.

Number of iterations

Number of iterations = number of passes, where each pass uses the number of samples [batch size]. To be clear, one pass = one forward pass + one backward pass (we don't count the forward pass and the backward pass as two different passes).

An epoch = one forward run and one backward run of all training examples

Batch size = number of training examples in a forward / backward pass. The larger the batch size, the more space you will need.

Example: In 1000 training examples and when the batch size is 500, it would take 2 iterations to complete 1 epoch.

After the text was transformed into vectors (numbers), then we applied this dataset to each algorithm. We set the training data size into 66 %. That means 36 % of the dataset will be used as test data. Figure 34 shows the result of this first experiment.

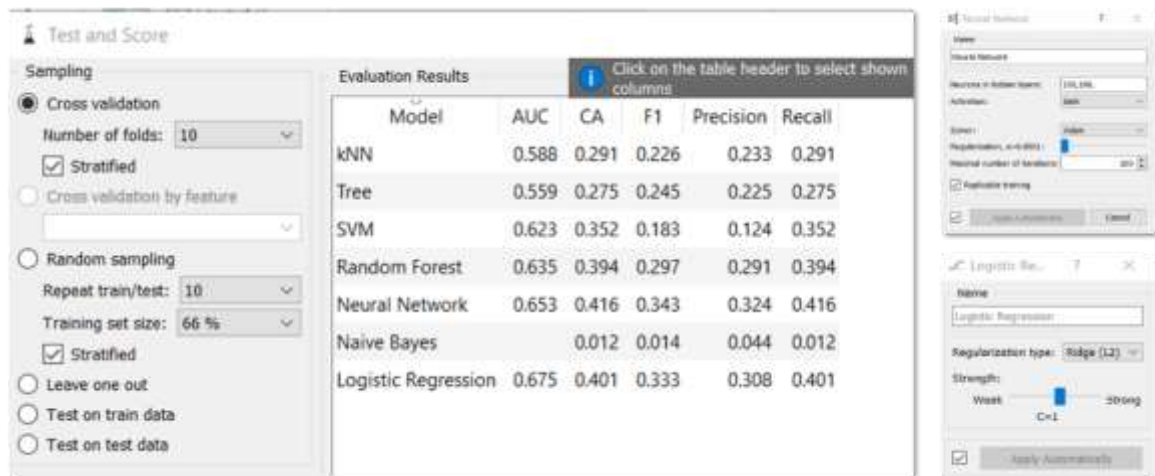


Figure 34: Comparison of machine learning algorithm performance

From the table shown in the figure 34 we can see the value AUC, which represent the performance of each algorithm. The performance means here the level of accuracy of data classification process. The best accuracy was reached by Logistic regression, AUC : 0,657. The second best was by Neural Network, AUC: 0,653. Based on this result we use Logistic

regression and Neural Network for the next step when we tried to optimize the performance by adjusting the parameters of each algorithm.

5.3 Concept for Image Analysis and Classification

The first step is as always, the selection of the appropriate tool for image analysis. In the concept of InnoCrowd two input types are analyzed in order to classify the new product. This classification can be based on input as text information or as image information and the integration of both input types. Basically, when there is only text file as the only input format, for example in the crowdsourcing project for the development of a software, then the classification of text is the only thing to be done.

5.3.1 Selecting the appropriate tool

We pre-selected 5 best available tool for the image analysis which are broadly used in industry as well as in research. They are namely OpenAI Gym, Google TensorFlow, Microsoft Cognitive Toolkit, Orange – Data Mining and PyTorch.

OpenAI Gym. The non-profit organization OpenAI is an initiative supported by Microsoft and Elon Musk, which aims to help develop artificial intelligence into a useful helper for humanity. Indeed, some great thinkers like Stephen Hawking have or had legitimate concerns that self-improving AI could quickly turn out to be a threat to humanity. To counteract a dystopian scenario, OpenAI supports the research and development of helpful artificial intelligence. AI developers can benefit from this idea: In addition to the idea of making as much AI technology available as open source, the OpenAI Gym is a simple toolkit for developing simple learning algorithms. In this way, developers can playfully familiarize themselves with how deep learning algorithms work. If you want more, you can also use the OpenAI API to deal with the language processing of AI systems; a look at the OpenAI beta channel is definitely worthwhile.

Google TensorFlow. Google also provides a free AI platform for developers: under the name TensorFlow, the search engine giant offers an open-source library in Python for AI development. As an "end-to-end open-source platform for machine learning" (Google self-promotion), TensorFlow allows beginners and professionals to use customized API

connections to deal with the development of AI systems. The name "tensor" comes from the arithmetic operations that take place on artificial neural networks. TensorFlow is already used internally by Google, which is why it is a very mature platform and supports numerous programming languages.

Microsoft Cognitive Toolkit. It is highly efficient and designed for scalability - this is the best way to describe the Microsoft Cognitive Toolkit, or CNTK for short. To ensure this, the framework is optimized for use on Microsoft's Azure. Its strengths lie primarily in the real-time analysis of data. Microsoft itself has long been using the toolkit in services such as Cortana or Skype, which means that the cognitive toolkit can be used wherever large amounts of data have to be analyzed.

Orange – Data Mining. For image analysis Orange can use deep network embedders. Technically, Orange would send the image to the server, where the server would push an image through a pre-trained deep neural network, like Google's Inception v3. Deep networks were most often trained with some special purpose in mind. Inception v3, for instance, can classify images into any of 1000 image classes.

PyTorch. PyTorch also originally came from one of the big players: Developed by Facebook engineers from the Torch environment that has existed since 2002, the open-source framework has meanwhile blossomed into one of the standard tools for AI development.

At first, we wanted to put Microsoft Cognitive Toolkit into a second priority, because the usage of it is related with a commercial tool of Microsoft Azure. If there are still enough choices of open-source tools with good quality, then we prefer the open-source tools. OpenAI Gym is also not quite optimum because it covers only simple learning algorithms. Google TensorFlow and PyTorch have becoming standard tools for image analysis, but they require programming skills. So, they are also not prioritized. The tool with good features and which is also user friendly is Orange – Data Mining. We prefer this tool because of the features for the data mining on one side. On the other side by using orange, we can apply an integrated data chain for text and image analysis. It is optimum in our work context.

5.3.2 Process chain for image analysis

We developed a process chain for the image analysis. Figure 35 shows the process chain. It starts with the widget for importing the images. The image is then transformed into numbers

by using the feature “Image Embedding”. We used 2 machine learning algorithms and compared their performance for the text classification task. The result of the comparison can be seen on the feature “Test and Score”.

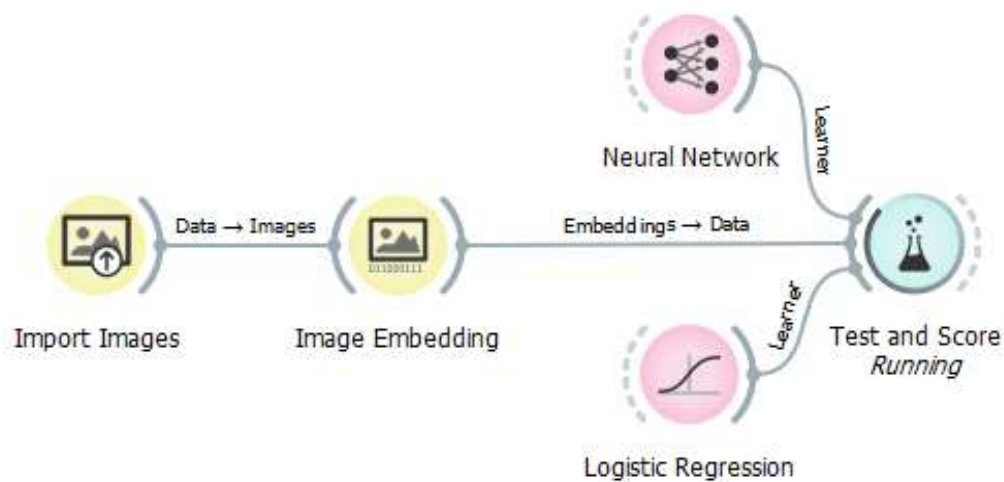


Figure 35: Process chain for image analysis

5.3.2.1 Import Images

Import Images walks through a directory and returns one row per located image. Column with image path is later used as an attribute for image visualization and embedding. It loads images and creates class values from folders. Columns include image name, path to image, width, height and image size. We can also load a folder containing subfolders. In this case Orange will consider each folder as a class value.

Data Table						
Info						
26 instances (no missing values)						
No features						
Discrete class with 2 values (no missing values)						
5 meta attributes (no missing values)						
Variables						
<input checked="" type="checkbox"/> Show variable labels (if present)						
<input type="checkbox"/> Visualize numeric values						
	category	image name	image da/Zecano/Import image	size	width	height
10	Monet	Tulip fields L...	Monet/Tulip ...	563108	2536	2044
11	Monet	Wild Popple...	Monet/Wild ...	356167	2536	1826
12	Monet	The Cliffs at ...	Monet/The C...	559355	1600	1279
13	Monet	Woman with ...	Monet/Wom...	2122240	1576	2399
14	Monet	Pont Argent...	Monet/Pont ...	338509	1600	1200
15	Monet	Music in the ...	Manet/Music...	2126219	3000	1906
16	Manet	Argenteuil	Manet/Argen...	229376	596	761
17	Manet	The Balcony	Manet/The B...	164649	655	799

Figure 36: Import Images

5.3.2.2 Image Embedding

Image Embedding reads images and uploads them to a remote server or evaluate them locally. Deep learning models are used to calculate a feature vector for each image. It returns an enhanced data table with additional columns (image descriptors).

Image Embedding offers several embedders, each trained for a specific task. Images are sent to a server or they are evaluated locally on the user's computer, where vectors representations are computed. SqueezeNet embedder offers a fast evaluation on users' computer which does not require an internet connection. If the user decides to use other embedders than SqueezeNet, he needs an internet connection. Images sent to the server are not stored anywhere.

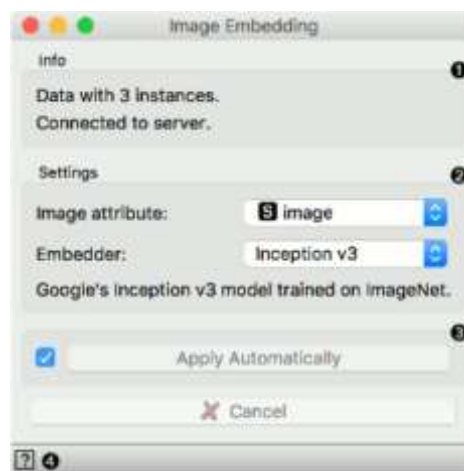


Figure 37: Image Embedding

Available embedders are InceptionV3, SqueezeNet, VGG16, VGG19 and DeepLoc.

InceptionV3 is Google's deep neural network for image recognition. It is trained on the ImageNet data set. The model we are using is available here. For the embedding, Orange uses the activations of the penultimate layer of the model, which represents images with vectors.

SqueezeNet is a deep model for image recognition that achieves AlexNet-level accuracy on ImageNet with 50x fewer parameters. The model is trained on the ImageNet dataset. Orange re-implemented the SqueezeNet by using weights from the author's pretrained model. It uses activations from pre-softmax (flatten10) layer as an embedding.

VGG16 and VGG19 are deep neural networks for image recognition proposed by Visual Geometry Group from the University of Oxford. They are trained on the ImageNet data set. Orange uses a community implementation of networks with original weights. As an embedding, it uses activations of the penultimate layer - fc7. Image Embedding also includes Painters, an embedder that was trained on 79,433 images of paintings by 1,584 painters and won Kaggle's Painter by Numbers competition. Activations of the penultimate layer of the network are used as an embedding.

DeepLoc is a convolutional network trained on 21,882 images of single cells that were manually assigned to one of 15 localization compartments. Orange uses the pre-trained network proposed by authors.

5.3.2.3 Machine Learning Algorithms for Image Analysis

Image analysis is a domain where mainly Neural Network gives the best data classification performance. But in context of this PhD work we considered also Logistic Regression as a qualified algorithm. We checked the both performance in order to get the best one.

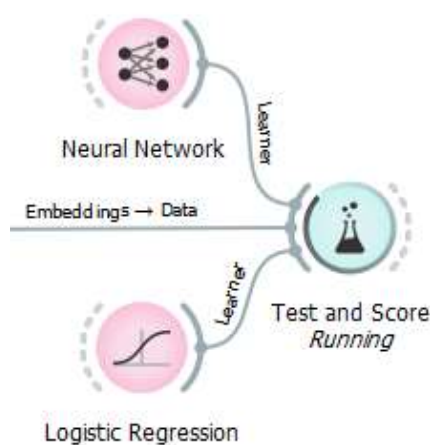


Figure 38: Machine Learning Algorithms for Image Analysis

This time we compared Neural Network and Logistic Regression. The neural network performed the best image classification, between 95% and 98% (fig. 6%).

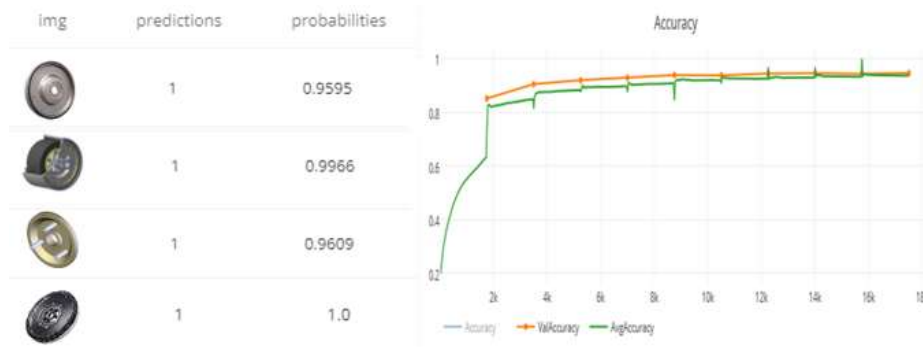


Figure 39: Result of the image classification with deep learning Neural Network

5.4 Integration of Text and Image Analysis

The third experiment is the integration of the text and the image processing (Figure 40). The process chain is integrated so that the system will give the final classification result. We used only the neural network, the best performer in previous experiments.

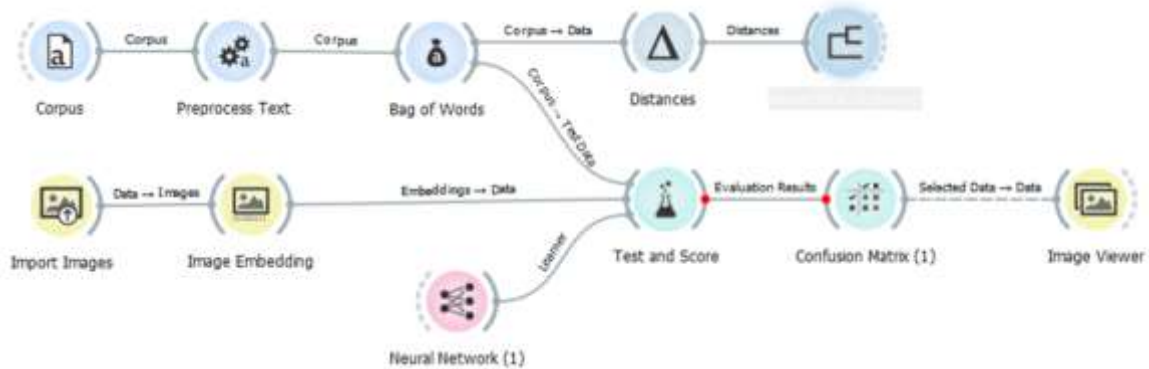


Figure 40: Process chain for the integrated text and image analysis

We carried out further experiments by manipulating each parameter. Roughly these are parameters that modified during the experiments:

KNN layer:

- number of layer: 3, 4 and 5
- neuron number in layer: 50, 60, 70, 80 90 100
- Regularization: 0,01 0,015 0,02 0,025
- Max. nr of iterations: 200, 230, 250, 270, 300, 330, 350

This experiment found the best result with the following parameters:

Neural network these are the parameters:

- Neuron in hidden layer: 80,90
- Activation: tanh
- Solver: Adam
- Regularization: 0,015
- Max. nr of iterations: 330
- Final best accuracy (AUC): 0,90

This 90 % accuracy is sufficient [Vallantin, 2018] to significantly supporting decisions on product specification / requirement, project time and cost, and the working / crowd team.

Chapter 6 SCALE-UP THE CONCEPT OF INNOCROWD

Although the final best accuracy reaches already 90%, it still necessary to optimize the neural network in order to:

- Upgrade the accuracy
- Shorten the processing time
- Simplify the network when possible in order to avoid overfitting

Overfitting means that the algorithm basically learns the data set "by heart", but does not recognize the underlying pattern or system. This means the algorithm is not able to classify new or unknown data. Overfitting is also common in data analysis. In this context, patterns and interrelationships are recognized that actually do not exist.

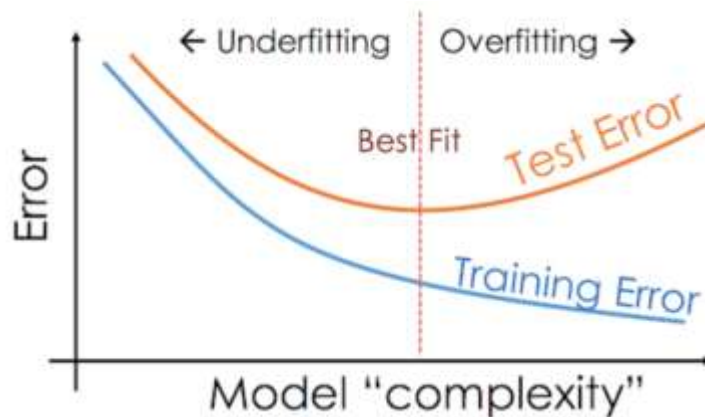


Figure 41: Overfitting and underfitting

The first approach for avoiding overfitting: more data. This definitely helps, but is not always feasible or associated with high costs (e.g. surveys). The second approach is a simplification. There are some approaches that force a simplification of the models and thus counteract overfitting. Overfitting neural network can be prevented by using the dropout technique. One part is randomly excluded from the algorithm. This influences all further connections of the

nodes. After repeated use, the individual models can then be reassembled to form a stable model.

Underfitting is the opposite of overfitting. It occurs when the network is too little adapted to the training examples.

This chapter deals with possible methods to optimize the performance and how we applied the related method in our solution.

6.1 Methods for Optimizing the Neural Network Architecture

The problem of Neural Network (NN) optimization is very difficult to grasp. Identifying a good result can often not be achieved with just one method. Therefore, different methods are being researched for handling this issue. The most promising methods for creating network architectures were divided into four groups:

- Intuitive method [Ali, 2019]
- Construction method [Sawalhi, 2014]
- Evolutionary method [Ding, 2013]
- Screening method [Lin, 2017]

6.1.1 Intuitive method

The search for a suitable network topology can certainly be a big challenge. Unfortunately, topology optimization is not only complex, but it is also mostly beyond human understanding. This is why 'network experts' are in demand who, with their experience, are able to not only to practice 'trial and error' methods but have the relevant knowledge and experience to model the network. The 'naive expert' certainly works less goal-oriented, but still comes to a useful result. This should be the most common procedures in application areas for NN topology optimization [Ali, 2019].

6.1.2 Construction methods

6.1.2.1 Deconstructive methods

A very large neural network (NN) can be optimized by deleting weights from them if they are believed to be of little importance on the net. This method is recommended if the NN can be trained well, but his generalization ability is not sufficient because there is insufficient training data [Sawalhi, 2014]. Definitely these methods are not particularly efficient: to choose excessively large networks and has to spend a lot of CPU time for training, but the weights are removed afterwards. So far this has been the method that is not really used but could be interesting in case of insufficient training data.

6.1.2.2 Constructive method

The constructive method is exactly the opposite. We extend the NN until it adequately depicts the problem. The construction process of the NN begins with a very simple structure, usually just one input and one output layer [Sawalhi, 2014]. This configuration is trained until the learning error stagnates. If the learning error is not less than a limit value, the network receives a first added hidden layer. The NN is trained again until the learning error stagnates again.

6.1.3 Evolutionary methods (Genetic Algorithm)

There is an approach to have rough optimization carried out by genetic algorithm [Ding, 2013]. An initial network is used to create a population with a number of epochs. In a multi-stage process, evolution mechanisms such as crossover, selection, mutation, among other things are used for network optimization. One of the main problems of genetic algorithm is the multitude of parameters that control the evolutionary process control or random numbers that lead to different results.

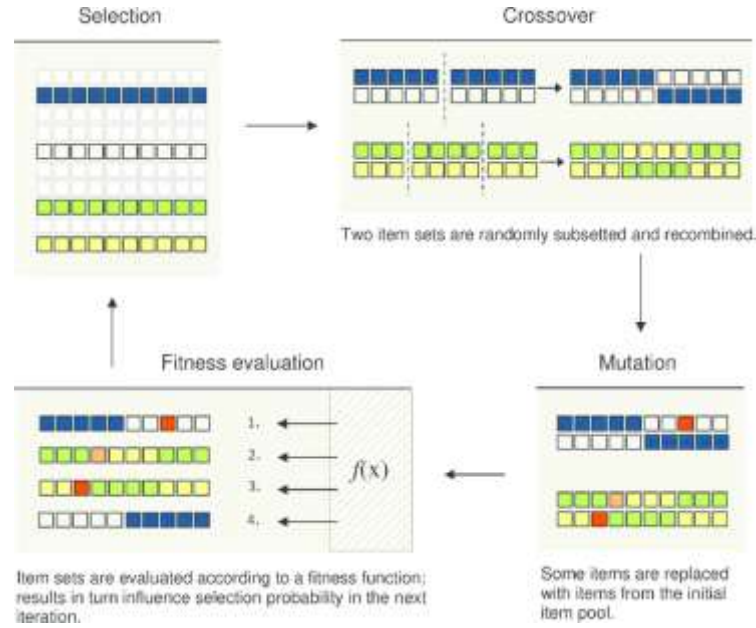


Figure 42: Genetic Algorithm

6.1.4 Screenings

Screening is a very good way to optimize NN. Both fine and coarse optimization can be implemented. During the fine tuning, test values can be used for the parameters of the learning rule, e.g. maximum, minimum and mean value of the recommended range of values [Lin, 2017]. The topology can be determined from a minimum to a maximum number of neurons. The screening appears to be inferior to the EA in the way that the space of the possible NN structures can only be examined discretely.

6.2 Optimization of the Neural Network of InnoCrowd with the Help of a Genetic Algorithm

This chapter explains the implementation of a genetic algorithm, which optimize the network topology of neural networks with any connections target.

The first section describes how the nets are represented in the genetic algorithm. Then the entire process of the algorithm is described and individual steps such as the generation of the starting population, the transformation between the display forms, the evaluation of fitness

and the crossover operator are detailed explained. Finally, the result of the final optimization is presented.

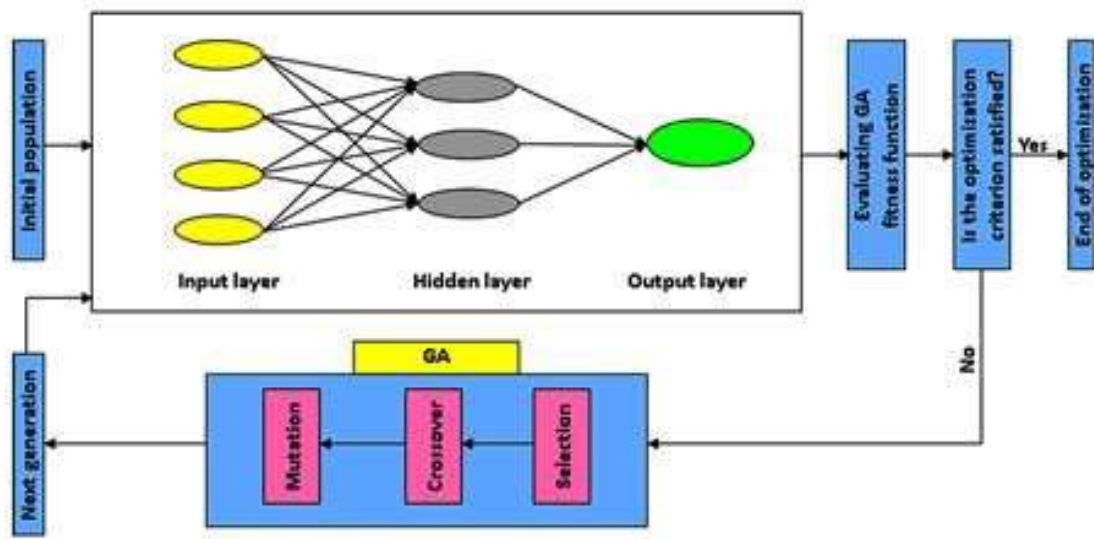


Figure 43: Optimization of the neural network architecture with the help of a genetic algorithm

6.2.1 Basic configuration of genetic algorithm

The aim of the genetic algorithm is to choose the connections among the neurons so that the training error is minimal. The algorithm should vary the number of neurons in the network. The number of input and output neurons is initially determined and is not changed because it is defined by the problem. In addition to the connections themselves, the number of connections in the network are also being optimized because each additional connection increases the training time significantly.

The genetic algorithm optimizes the individuals on the basis of the genotypes, the representation of the individuals in the genetic algorithm is therefore a central aspect. The form of representation of the individuals must be chosen so that the parameters to be optimized are represented and the algorithm can optimize with the help of genetic operators. In the genotype it is indicated how many neurons there are in which layer and which activation function is used in each neuron. The activation function is assigned at the level of the layer

types. So, there is an activation function for all input neurons, one for all hidden neurons and one for all output neurons. The other part of the genotype consists of the learning function used in training the network and the associated parameters, which vary depending on the function.

The phenotype is a completely initialized neural network that contains the information from the genotype. It should be explicitly pointed out here that the weights between the neurons is not part of the optimization. The data of the genotype, which serves as the basis for the phenotype, is always interpreted as a fully connected feedforward network without shortcut connections, which means that all neurons of layer n (seen from the input layer) with all neurons of layer $n + 1$ are connected. There are no connections from the output layer. There are no connections from the back to the front layer and no connections within a layer.

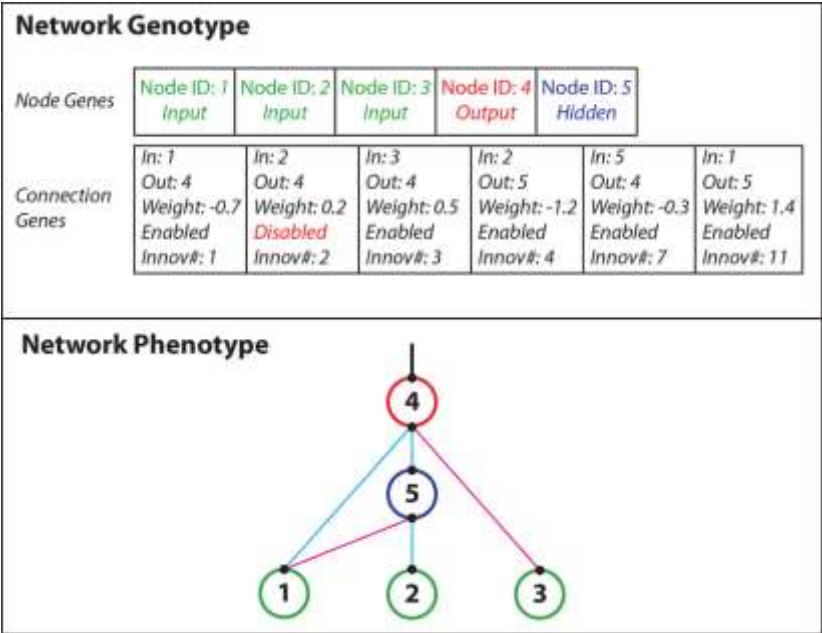


Figure 44: Genotype and Phenotype

6.2.2 Implementation of the genetic algorithm for InnoCrowd

Each network is trained until one of the following events occurs: 1. The desired square error (which decreases slightly per generation) is not exceeded. 2. The number of training cycles is greater than the number of training cycles of the 16th worst individual from the previous

generation multiplied by a factor $\times 4$ th. After the run, the fitness is determined for each individual. The 16 networks with the best fitness are selected for reproduction for the next generation, which corresponds to the Elitism replacement scheme.

Exactly one mutant is generated from each selected network. This does not necessarily have to differ from its predecessor. To determine if and what during the mutation is changed, the mutation rate is used. This is 0.3 for all experiments. That means there is a 30% probability of a change for each mutable characteristic.

The actual networks are mutated at the hidden layer level. Every hidden Layer in the network is checked against the probability of mutation to see whether there will be a change. If there is a change, a random value is determined and the number of neurons in the relevant layer is set to the determined value.

6.2.3 The individual steps for the optimization of the network topology

The basic sequence of a genetic algorithm has already been dealt. The individual steps of the algorithm will now be explained in relation to the optimization of the network topology. In the first step, a starting population is created, whereby the individuals are represented in the matrix. Then all individuals are brought into the form, which stores four arrays for all start indices, destination indices and weight indices, as well as the value of the delay of a connection. Algorithm can thus be used again for training the networks. Both forms of representation of the individuals are saved so that the conversion only has to be carried out once per individual. The training provides a mistake for each individual, with the help of which the networks are evaluated. The more precise calculation of the fitness function is explained. On the basis of fitness, the parents are selected who will produce new offspring after using the genetic operators. The offspring are also converted into the “network representation” and trained and can then be integrated into the population.

6.2.4 The best NN parameter for InnoCrowd

For the method presented, all generation quantities come with the integer divisible through six. In this work only a generation size by forty-eight individuals is used. First, forty-eight individuals are randomly generated. The activation functions of the individual layers, the

learning function, the bias of the individual neurons and the weighting of the connections between the neurons are determined per network randomly, these collected networks are transferred with appropriate control instructions to Orange. Orange trains the neural networks based on the given training pattern. As a control, the simulator checks after each learning cycle with the help of the validation pattern the generalization ability of the respective network, which happens here via the quadratic error. The termination criterion is the achievement of a given square error or the achievement of a certain number of learning cycles. Both termination criteria vary from generation to generation. After the training, the network simulator returns the squares achieved and the learning cycles required for all networks. These values and the size of the neuronal networks are incorporated into the fitness function, which is used to rank the individuals. These are the parameters of the neural network after the final optimization through genetic algorithm:

- Neuron in hidden layer: 55, 70
- Activation: tanh
- Solver: Adam
- Regularization: 0,013
- Max. nr of iterations: 260
- Final best accuracy (AUC): 0,95

According to Vallantin [Vallantin, 2018] the accuracy between 80% and 90% is excellent. That means 95 % is more than sufficient to significantly supporting decisions on product specification / requirement, project time and cost, and the working / crowd team.

Chapter 7 INNOCROWD APPLICATION IN 3 REAL LIFE CASES

This chapter shows the application of InnoCrowd in the industrial use cases. Three real life use cases are shown and discussed here. One use case is the most common use case in the industry, where a new product design is needed by the crowdsourcer. The second use case is the application of InnoCrowd in the area of Information Technology, where the user wants to get an implementation of a specified software. And the third use case is in the area of architectural design which can be seen as one of the most applied use case in the crowdsourcing project [Sawalhi, 2014].

7.1 Application in Context of Product Design

Crowdsourcing task related to product design is very common. We choose one example where a user, the crowdsourcer, wanted to have a design of a drone. He defined a rough product design and also given how the product may look like. This information is taken as input data. The input data has two forms, in text and in image form. We provided this input data to the engine of InnoCrowd and let it classify. At the end the user will get result in form of founded related crowdsourcing projects.

These items belong to Product Design under CADCROWD:

- 3D Product Demo
- Concept Design
- Custom Cast Urethane Molding
- Design for Assembly Services
- Exercise Equipment Design
- Industrial Design Services

- Internet of Things (IoT)
- Lighting Fixtures Design
- Medical Devices
- Open Innovation
- Packaging Concept Design
- Packaging Design
- Product Engineering Services
- Prototypes for Shoes, Sneakers, and Athletic Runners
- Rapid Prototyping Services
- Tool Design Services

7.1.1 Input data

The crowdsourcer wanted to have a design of a racing drone. All the requirement and images are taken from the crowdsourcing platform CADCROWD. The initial product requirement is the following.

Title:

Racing drone (quadcopter) from notes to CAD and 3D model

Specification:

- The drone will consist of two pieces of moulded carbon fiber:
 - bottom hull
 - hood
- Key features of the drone:
 - two piece construction with interlocking lip
 - tube arms (16mm outer diameter, 12mm inner diameter)
 - fit 6" diameter propeller (152.4mm)
 - 16x19mm standard motor mounting
 - fit standard HS1177 camera with adjustable camera angle from 0° - 70° of tilt
 - fit standard flight controller circuit board (30.5mm square mounting holes on the bottom)

- 4 mm thick at motor mount
 - 3-4 mm thickness of hull
- Deliverables:
 - Technical CAD model of drone components
 - Inverse CAD model for the moulds for both pieces
 - 3D renders
- Bonus points for:
 - finite element analysis to highlight structural weaknesses and suggest improvements
 - include motors, propeller, and other electronics in the 3D renders
- Looking for:
 - Accuracy
 - aggressive style
 - Prefer a sketchup model.

The figure 45 shows image files given by the crowdsourcer, which represents the possible design of the product.



Figure 45: Sample of the product picture

7.1.2 Classification result done by InnoCrowd

The input data as text and image files were fed into the engine of InnoCrowd. It classified the data and given the related past crowdsource projects.

Classification result:

- It took 165 seconds for InnoCrowd to give the classification results
- Three past projects were relevant to this input data with matching rate of 93%, 86% and 80%

The first related past project with the matching rate of 93 % is shown in figure below.

The screenshot shows a crowdsource project page for a drone. The title is "Simple Miniature Racing Drone for STEM / STEAM Learning". It is categorized under "Toys & Games" and "STEM / STEAM TOYS". The project is in the "Ended" state. The description mentions that the drone should be simple, field serviceable, and designed for kids. It lists requirements for the body, shape, battery, and flight characteristics. The project has 10 entries, 6 comments, and 3 attachments. The prize is \$800 for the first place and \$400 for the second place. The project was created on 9 Oct 2018 and ended on 20 Oct 2018. The attachments include a PDF file named "PRO-DRONE-RACE - Design Contest".

Simple Miniature Racing Drone for STEM / STEAM Learning

in Toys & Games held by STEM / STEAM TOYS · Last seen: 2018-10-10 21:33:19

Open / Choosing Finalist / Ended

Description Entries 10 Comments 6 Attachments 3

Description:

We have a few STEM / STEAM (science and tech courses for kids) programs we are designing for, including a miniature radio controlled racing drone. They should run off of a single rechargeable A or AA battery or rechargeable easily found (commonly available) battery. Need designs for the computer controllers (Arduino or similar), software, simple radio remote control and system, electric motor systems, and drive systems engineered. We have the basic frame shape. We are looking for simple, field serviceable designs that kids can take apart and put together and use to learn these technologies. Just like everyone, we want it ASAP.

Wants:

Size: Body about 8-10 CM, 3-4 inches long. Motor struts can project out from this shape.
Shape: Based on our logo shape, attached as a JPG. Two basic plates (upper and lower plates) can be 3D printed or cut from carbon fiber sheets.
Battery: A single A (7 huc) or AA (5 huc) rechargeable, or other commonly available battery.
Drone: 4 rotors, attachable prop guard, stable flight characteristics (so will need sensors and feedback system to dampen instability).
Basic flight commands and control characteristics:
1 - Take off slowly to 1 meter high by simple radio control button.
2 - Land slowly straight down by simple radio control button.
3 - Forward and back by simple radio control.

Payment Pending

Prize #1: \$800
Prize #2: \$400

Project Info

Created: 9 Oct 2018
Ended: 20 Oct 2018
Entries: 10
Designers: 7

[Report Contest](#)

Attachments

PRO-DRONE-RACE - Design Contest

Figure 46: The first related project to “Drone”

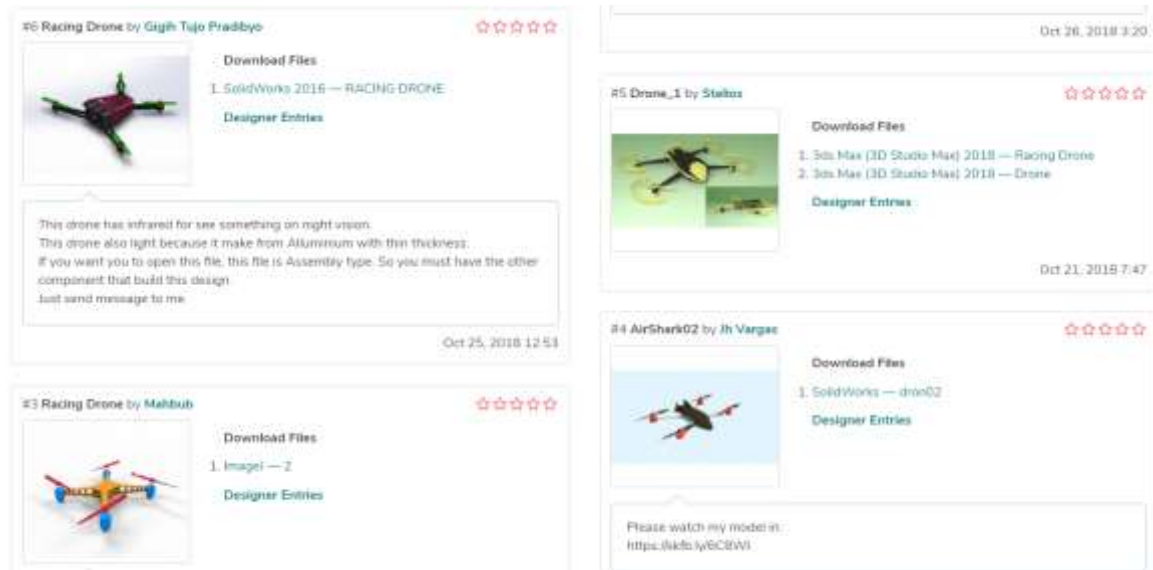


Figure 47: Images to the first related project to “Drone”

The second related past project with the matching rate of 86 % is shown in figure below.

Transport UAS/UAV/Drone

in 3D Modeling Design held by UASCA - Last seen: 2016-08-02 03:12:33

Open | Choosing Finalist | Ended

Like | Share | Tweet | Mark

Post Similar Contest

Description | Entries (92) | Comments (32) | Attachments (3)

Description:

The purpose of this project is to look into the feasibility of creating a proof of concept electric powered UAV that is autonomous, can carry varied payloads (fixed camera, camera on a gimbal, carry case to transport goods, etc.), has a flight time of 60+ minutes, and most important, radical in design! Must have at least 4 engines. Not looking for a COTS solution or similar. I would like to have the UAS to be modular for various mission capabilities: Mission (Crime Prevention, Fire Management, Emergency Response, Intelligence, Fire Investigation & Damage Assessment, Accident Reconstruction, Search and Rescue, Forensics, Event & VIP Security, Tactical situational awareness & targeting, Perimeter & convoy security, Covert Intelligence, Surveillance and Reconnaissance, Anti-piracy, tactical ship boarding, De-mining, Emergency & disaster response HAZMAT/CBRNE management, etc.) payload(s) and flight characteristics might include:

- Surveillance with Gimbal, or Gimbals to allow for 360 degree view
- FLIR, HD RGB (or HDMI if possible), NIR, NDVI, LWIR, Lidar
- Area Mapping for mitigation, disaster recovery, assessments, inspections, etc.
- High Speed with Heavy Lift capability for emergency medical devices/medications.
- This is for cargo, with an understanding that a balance will need to be struck to determine maximum flight time.
- Tether Mode (for training and long term situational awareness).
- The possibility of powering unit and payload feed via tether.
- Modular, hot-swappable, payload chamber(s) for Sensor Arrays for various scenarios that might include:
 - o Collision Avoidance
 - o Acoustic Detection and Location (Military, Public Services)
 - o Electrochemical and Infrared Imaging Gas & Liquid Detectors (Environmental Air Sensing, Gas and Liquid Leak Detection, CBRNE)

Prepaid Prize

Prize #1: **\$1,200**

Project Info

Created:	2 Jun 2015
Ended:	22 Jun 2015
Entries:	92
Designers:	19

Report Contest

Attachments

- 5fup3
- qual-vtni concept by ferain-d6e7xz
- Schema bell qtr

Figure 48: The second related project to “Drone”

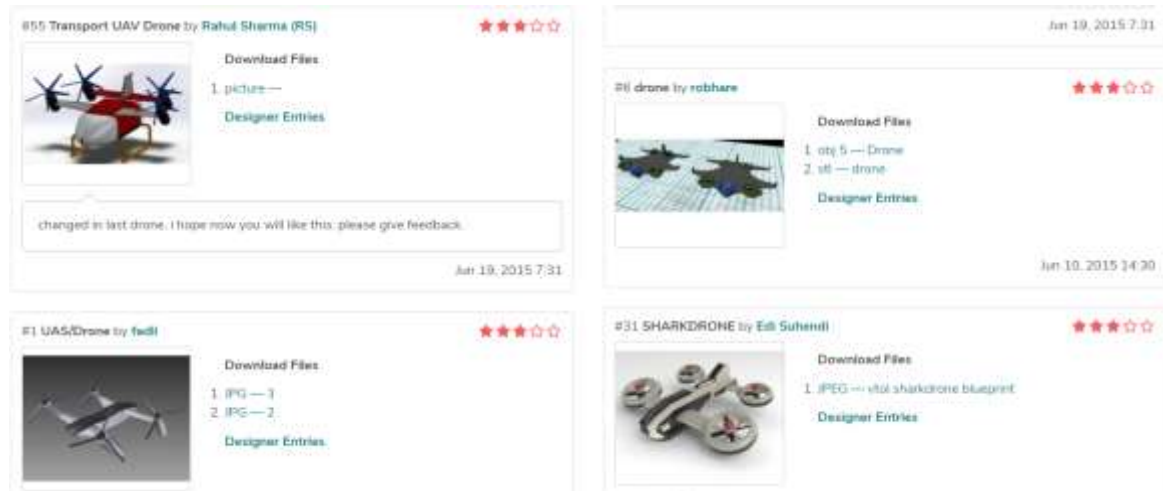


Figure 49: Images to the second related project to “Drone”

The third related past project with the matching rate of 80 % is shown in figure below.

Personal Light Drone Steady

in 3D Modeling Design held by gilbertoo · Last seen: 2020-11-13 00:19:38

Open · Choosing Finalist · Ended

Like 0 · Share 0 · Tweet · Monitor

[Post Similar Contest](#)

Description · Entries **50** · Comments **14** · Attachments **4**

Description:
Looking for someone to translate drawings into 3D models. Product is a personal drone that emits light, launched from a cylindrical housing unit.

Wants:
Essentially, we're looking for something that will be useful in a concept presentation. Colors differentiating different parts of the product would be great, to better clarify how everything fits together.

Additional Information
Measurements:
4-5 in diameter for the drone
Drone no taller than 2 in
About 5.25 in diameter for the cylindrical case
11 in length for the case

Prepaid Prize
Prize #1: **\$200**

Project Info

Created:	12 Oct 2020
Ends:	11 Nov 2020
Entries:	50
Designers:	36

[Report Contest](#)

Logo Attachments

[Torch light-01](#)

Attachments

[PXL_20201008_185722752](#)
[PXL_20201008_185720008](#)

Figure 50: The third related project to “Drone”

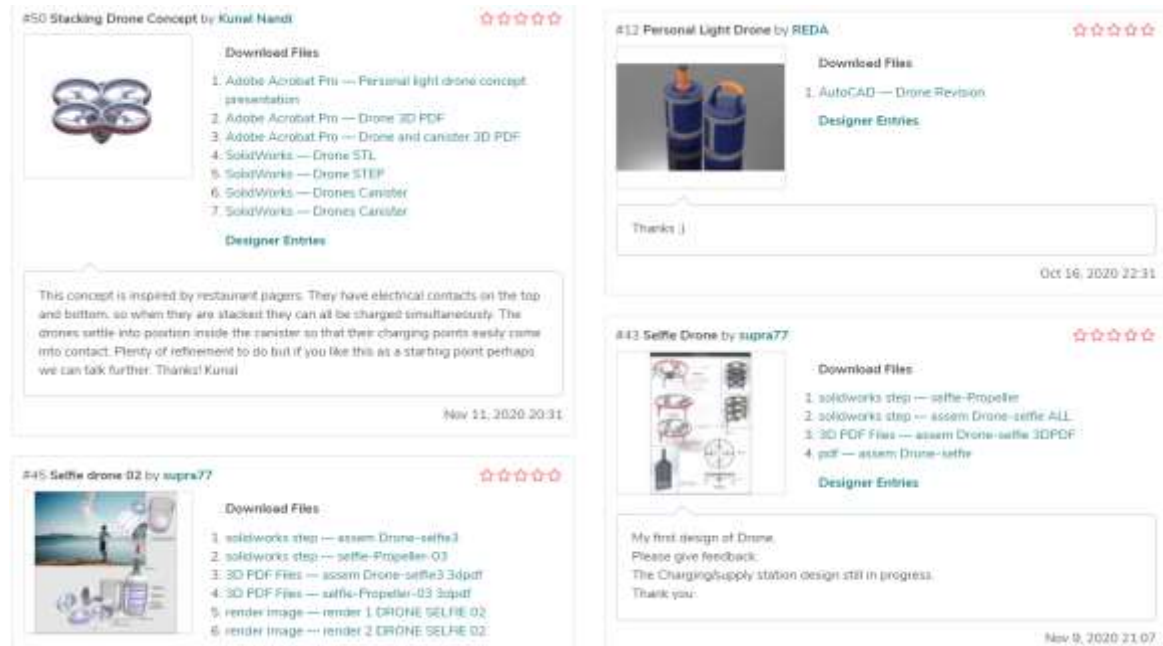


Figure 51: Images to the third related project to “Drone”

From the figure above the crowdsourcer can find not only the relevant product images but also the name of persons (the crowd) who have given their contribution. This information regarding the past working team is also important for measuring the project risk. If we know there are experts available in the community then we can be sure that will a good result produced by the crowd.

7.1.3 Discussion on the result

The classification result is more than just matching the words on title level. It is not like searching in google with a simple sentence as an input. It is much more than that because the algorithm bag of words and word phrasing find matching context in deeper level.

The results give three alternatives with each alternative having its own matching level. It gives a good overview about the related crowdsourcing project to the user and it will enable the user to decide the setup of his new crowdsourcing project based on the information given in the previous projects.

Three figures shown on the previous sub chapter give important information to the crowdsourcer regarding the following:

- Product requirement
- The main project cost in form of the prize for the winner
- The project timeline
- How much contributions from the crowd
- How much designer have contributed and also who exactly they are (information of their name and profile is accessible)

The reason of the best matching with 93 % matching rate are the following:

- The both crowdsourcing projects are similar in the project title: Racing Drone.
- The both product requirements have big similarities.
- Also images from both crowdsourcing projects are having similarities.

The reasons of the second best matching with 86 % matching rate are the following:

- The both crowdsourcing projects are asking for at least the same main product: Drone.
- The product requirements have similarities on some aspects: camera and 4 rotors.
- Images from both crowdsourcing projects are having similarities.

The reason of the third best matching with 80 % matching rate are the following:

- Both crowdsourcing projects are asking for at least the same main product: Drone.
- The product requirement on the past crowdsourcing project is very short, therefore it is harder to find matching bag of word vectors.
- Images from both crowdsourcing projects are having similarities.

By using all these three similar crowdsourcing pas projects as best practices / references the new user can decide the best setup as the following:

- We can see by comparing all three alternatives that the bigger the prize the higher number of participating crowds. 800 dollars is acceptable, but 1200 dollar is off course very much more interesting. Depend on the project budget and on the complexity of the product, the new user can decide between 800 dollars until 1200 dollar.
- We can see by comparing all three alternatives that the longer the project timeline the higher number of participating crowds. 20 days are good, but 30 days are better. The better decision might be 30 days.
- Short description on product requirement may make the crowd think that it is an easy task, so that many people tried to contribute. But it might be problematic on the end product quality, if it fulfills really the needed product specification.

7.2 Application in Context of Information Technology Task

7.2.1 Input data

The crowdsourcer wanted to have a specific website design. All the requirements are taken from the crowdsourcing platform TOPCODER. The initial product requirement is the following.

Title:

Odyssey Corporate Website Home Page Challenge

Specification:

The application will connect to a Drupal backend to get its content. In this challenge we want to build the Home Page.

We believe that our customers have more important things to do than wrestle with complex products or spend valuable time administering their banking arrangements. We recognize that

professional and personal lives often overlap and our mission is to help empower our customers to achieve their ambitions by serving all their banking needs.

Banking customers are looking for a bank that can help them make their deal happen in an efficient, secure, fast, and flexible way as much as it can.

We understand different people want different things from "Relationship management"

We expect a countertrend of "committed" relationships as customers react to digitization in our services.

Our goal in this project is to create strong direct relationships with our customers. We want to provide everything - a blended bank, preferential banking for their family, and open to non-banking services via our digital app.

Technology Stack

- React.Js 17.x
- Typescript
- CSS
- HTML5

Individual requirements

- Create a brand new React.Js application
- Initial application will be provided in the forum
- Use best practices for structuring the application (<https://reactjs.org/>)
- The react components will be mapped into a Drupal content or paragraph, therefore all frontend data will be fetched from Drupal endpoints (see private section)
- Only the home page is in scope. Marvel app link in the private section
- Top header (black background) is static, meaning it won't load from Drupal
- Secondary menu is dynamic and should be loaded from the Drupal endpoint - please use the appropriate filter
- Footer will also come from Drupal
- Lowest footer is static, meaning it won't load from Drupal
- The home page will need loading several different dynamic content. The initial content URL will be provided in the private section
- Make sure to follow the margin, padding, font sizes in Marvel app

7.2.2 Classification result done by InnoCrowd

Classification result:

- It took 45 seconds for InnoCrowd to give the classification results. Shorter time because this time only text classification (without image classification)
- Three past projects were relevant to this input data with matching rate of 95%, 88% and 76%

The third related past project with the matching rate of 95 % is shown in figure below.

The screenshot displays a challenge page with the following sections:

- Challenge Details:**

The main goal of this challenge is to design a clean and easy to user experience (UX) for the Financial Loans website application.

Our client offers loans to businesses that can't get a traditional bank loan, and unlike banks - our client gets the loan repayment(s) directly from his client's sales. This application will be used to provide an easy-to-use experience for new and potential customers applying for loans.

The goal is for the customer to self-service and to fill out the loan they need. We need you to focus on simplifying the process, focusing on making it easy and enticing for all new / potential customers. Customers will need to find out more details about getting a loan, get a loan quote, and apply for a new loan.
- Challenge Downloads:**

Use the provided wireframe content from the suggested application screens to help you get started, keeping in mind that we are looking for a clean design solution with minimalist look, focusing on the customer user experience and ease of use.
- Required Pages:**
 - **Home page** - will focus on the key message(s) for the new / potential customers, along with means to learn more about the product(s), get a quote and apply for a loan. New users should be able to register, and existing customers to log in to their account.

NOTE: we are not a bank and do not want to look like a bank. The goal is to quickly capture the user's information / request.

 - **Advance page** - explains the benefits of getting cash advancing from our client if they need more money, and presents the users with testimonials and options to compare the sample product with others and get a rate quote.
- Fonts, Stock Photos, and Icons:**

All fonts, stock photos, and icons within your design must be declared when you submit. DO NOT include any 3rd party files in your submission or source files. Read about the [policy](#)
- Screening:**

All submissions are screened for eligibility before the challenge holder picks winners. Don't let your hard work go to waste. Learn more about how to [pass screening](#).
- Questions? Ask in the Challenge Discussion Forums.**
- SOURCE FILES:**
 - Layered PSD files created in Adobe Photoshop or similar
 - AI files created in Adobe Illustrator or similar

You must include all source files with your submission.
- SUBMISSION LIMIT:**

Unlimited
- CHALLENGE TERMS:**

[Standard Terms for TopCoder](#)

Figure 52: The first related project to “Website Design”

The third related past project with the matching rate of 88 % is shown in figure below.

Welcome to **SiWC Brewery Wix Website Challenge**! This is the next in a series of challenges to build a website for SiWC brewery. We've completed the design phase and we are moving on to the development phase of the project. In this challenge, we need your help to convert the indicated screens into working web pages using the website creator tool Wix. Wix is a WYSIWYG (mostly) website building platform. There is no heavy coding required for this challenge. Some knowledge of JavaScript may be helpful. Read the challenge specification carefully and watch the forums for any questions or feedback concerning this challenge. Let us know if you have any questions in the challenge forum! # **TECHNOLOGIES** Wix Javascript # **GENERAL REQUIREMENTS** **Layout Specification** We are creating for desktop only, for this challenge. **Icon and image quality** Please use the images from the Figma prototype only. Objects included in the Wix templates will need to be updated to the matching asset. **Typography** Please use Barlow Medium for the main font. It is available in Wix. # **INDIVIDUAL REQUIREMENTS** We want to use the attached assets to create some website pages using the Wix platform. The final project will be a fully functional ecommerce site for the brewery. Future challenges will add the filter, search and other functions/elements. You will need to be added to the Figma project that holds the assets for this competition. A request thread is in the forum. Descriptions of screens and related functionality are below. **Landing Page (Desktop)** **Landing.png** – This is the main screen for the online store. – Clicking on **About** and **Shop** in the header should navigate to those pages. The other links are out of scope. – Clicking the **BUY NOW** button should direct the user to the **SHOP** page. – The **DISCOVER MORE** button should direct the user to the **Beer Details** page. – Please create the **FIND US NOW!** as a clickable element. We will link this to the appropriate page later. – **Join the Pack** section please use an available Wix plugin to configure the subscription service. **Shop Page (Desktop)** **SHOP.png** – The filter function is out of scope for this challenge. Please download the filter .png and use it as a placeholder in the appropriate area. – **Add to Cart** buttons should be separate elements. – The **DISCOVER MORE** button should direct the user to the details page. – Please configure the **Recently Viewed** section with one of the available Wix plugins. **Beer Details Screen (Desktop)** **Beer Detail.png** – The details should be left editable. – The **Item Added to Cart** button is not in scope. – The **Similar Beers** section

2021 Topcoder(R) Open

REVIEW STYLE:

Final Review:
Community Review Board

Approval:
User Sign-Off

CHALLENGE LINKS:
Review Scorecard
Useable Code Rules

CHALLENGE TERMS:
Standard Terms for Topcoder Competitions v3

TOOLBOX:
Topcoder Extension for VSCode
Topcoder Templates repository

SHARE:
f t b e +

BY SOFTWARE

Figure 53: The second related project to “Website Design”

The third related past project with the matching rate of 76 % is shown in figure below.

Challenge Summary

Welcome to the TCO22 Website Design Challenge.

Can you believe that it is almost the time? Stage 4 of TCO21 qualification period is almost over, so we are already preparing for TCO22. We would like you to come up with a great design for the TCO22 website based on the requirements put together by the Topcoder team.

CHALLENGE OBJECTIVES

- To create the UI of 4 screens in desktop and mobile format.
- To provide an engaging way to introduce the event and the journey to it.
- To explore improvements from the previous years' TCO sites.

ROUND 1

Please provide the following screens for checkpoint along with your clickable Marvel prototype with hotspots.

- 1. Homepage - Desktop
- 2. Competition Overview - Desktop
- 3. UI Design Competition Details - Desktop
- 4. Regional Event - Desktop

ROUND 2

Please provide the following screens with checkpoint feedback incorporated for final review along with your clickable Marvel prototype with hotspots.

- 1. Homepage - Desktop

ELIGIBLE EVENTS:

2021 Topcoder[®] Open

CHALLENGE LINKS:

[Screening Scorecard](#)

SUBMISSION FORMAT:

Your Design Files:

1. Look for instructions in this challenge regarding what files to provide.
2. Place your submission files into a "Submission.zip" file.
3. Place all of your source files into a "Source.zip" file.
4. Declare your fonts, stock photos, and icons in a "Declaration.txt" file.
5. Create a JPG preview file.
6. Place the 4 files you just created into a single zip file. This will be what you upload.

Trouble formatting your submission or want to learn more? [Read the FAQ.](#)

Fonts, Stock Photos, and Icons:

All fonts, stock photos, and icons within your design must be declared when you submit. DO NOT include any 3rd party files in your

Figure 54: The third related project to “Website Design”

7.2.3 Discussion on the result

The reason of the best matching with 95 % matching rate are the following:

- Both crowdsourcing projects are the same in the main topic: website design in finance institution
- Both product requirements have very big similarities in terms they used
- It seems like the InnoCrowd engine reaches its peak performance of 95% (as seen and reached during the training phase)

The reason of the second-best matching with 88 % matching rate are the following:

- The both crowdsourcing projects are asking for at least the same main product: website challenge.
- The product requirements on the past crowdsourcing project are long and detailed. It gives more change to more matching on bag of word vectors.

The reason of the third best matching with 76 % matching rate are the following:

- The both crowdsourcing projects are asking for at least the same main product: website challenge.
- The product requirement on the past crowdsourcing project having slightly different detail and therefore also different bag of word vectors

7.3 Application in Context of Architectural Design

These items belong to architectural design under CADCROWD:

- 2D Drawings and Floor Plans
- 3D AR/VR Architectural Services
- 3D Architectural Animation Services
- 3D Architectural Visualization Services
- 3D Visualizer Services
- ArchiCAD Design Services
- Architectural 3D Modeling
- Architectural BIM Services
- Architectural Detailing Services
- Architectural Drafting Services
- Architectural Drawing Services
- Architectural Model Making
- Architectural Planning and Design
- Architectural Presentation Services
- Architectural Site Plans
- Architectural Sketch Design Services

7.3.1 Input data

The crowdsourcer wanted to have a new design of a floor plan. The initial product requirement is the following.

Title:

Floor plan for remodel of home drop zone, gourmet kitchen, powder room, pantry

Specification:

Looking to remodel the kitchen with a large rectangular island, add a walk-in pantry, add a large drop zone from garage entry, and probably move powder room. Basement is unfinished so plumbing can be moved. Will minimally bump into the garage not to impede on parked car space. Would like powder room to be small like the attached example. Kitchen appliances to be 48" range top, double wall ovens, and 48-inch counter depth fridge. Possibly a second prep sink if room permits. Laundry room can be reconfigured and hall closet removed, but need to leave plenty of counter/working space in laundry room. Duct work runs up next to steps but only takes up small part of allotted space, even smaller than the second level duct plan reveals. Two-dimensional floor plan is adequate, 3D not necessary. The island should be at least 9 ft by 4 feet and the main sink placed in the island. Range should be the centerpiece in a bank of cabinet space. The space into the pantry can be a doorway with rear cabinetry, making it look like an entrance to another room. Take up as little garage space as is needed to complete the objective, still want room for tools, lawnmower, etc.

Don't bump into the hearth room, great room, or dining room space. Also don't move steps. The dinette needs to stay approximately the size it currently is.

The figure 55 shows image files given by the crowdsourcer, which represents the possible design of the floor plan.



Figure 55: Possible design of Floor Plan

7.3.2 Classification result done by InnoCrowd

Classification result:

- It took 327 seconds for InnoCrowd to give the classification results
- Three past projects were relevant to this input data with matching rate of 91%, 87% and 85%

The first related past project with the matching rate of 91 % is shown in figure below.

Improve upon our floor plan and hardscape design for a 4,500 sqft waterfront home Blind

in Architectural Design held by wilson123 - Last seen: 2019-11-04 01:57:11

Like 0 Store 0 Tweet 0 Mark 0

Post Similar Contest

Open Choosing Finalist Ended

Description Entries 0 Comments 11 Attachments 4

Description:
We have a plot of canal front land in the Cayman Islands. Looking to build a home on it soon, so considering floor plans. I attach a floorplan of our current ideas to give an idea of what we're thinking. I'm sure it can be improved upon, and we can make more of the space. Also attached is a brochure of the land to give an idea of the area. Ours is plot 17, so you can see the dimensions and orientation of the land.

I'd like to keep the total floor size (for the house, garage and apartment on top) to approx 4,500 sq ft.

Wants:
Required specs:
- 4 bedrooms (all upstairs)
- 2 ensuite bathrooms and 1 shared family bathroom for the other two rooms
- Master bedroom to be quite large and have a decent size walk in wardrobe, and large master ensuite bathroom (spa feel), preferably with double shower and separate room for toilet.
- An office, preferably with view over pool/canal
- Open plan large kitchen/dining/living area
- Separate living/play room
- Preferably a gym on the ground floor.

Prepaid Prize
Prize #1: \$400

Project Info

Created:	30 Sep 2019
Ended:	30 Oct 2019
Entries:	5
Designers:	4

[Report Contest](#)

Attachments

- 03 UPPER FLOOR PLAN 2
- 02 MAIN FLOOR PLAN 4
- 01 A HARDSCAPE 1
- Harbour Reach Brochure Crighton Mar

Figure 56: The first related project to “Floor Plan”

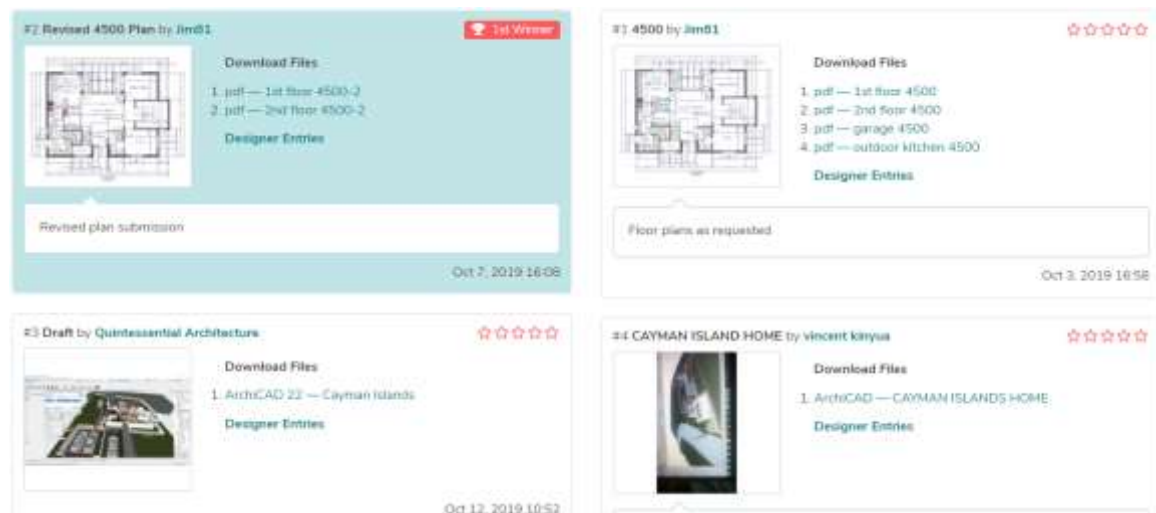


Figure 57: Images to the first related project to “Floor Plan”

The second related past project with the matching rate of 87 % is shown in figure below.

Looking for floorplan for "Duplex" Office Building (New Construction) Builds

in Architectural Design held by ASJT - Last seen: 2020-06-09 18:43:52

Open Choosing Finalist Ended

Like 0 Share 0 Tweet 0 Marken

Post Similar Contest

Prepaid Prize
Prize #1: **\$240**

Project Info

Created:	2 Jun 2020
Ended:	17 Jun 2020
Entries:	1
Designers:	1

[Report Contest](#)

Attachments

- Notes about office space
- Exterior duplex 1
- Model Exterior

Description:

We've purchased a piece of land and are building office space for two law firms, and are looking at potentially having a "duplex" model where we put two single-story spaces together, with separate entrances. We need some help working through potential floor plan layouts for each side of the duplex. These layouts will help us with our planning process for the site. We're looking for a floorplan design on how we can lay the spaces out in the most flexible way while still having the required spaces/offices that we mention in the attachments/notes. If there is "extra" space in a particular side of the building, we'd love your suggestions on how to fill it.

Wants:

We'll be looking for:

- maximizing space (getting "bang for our buck" in terms of layout/footprint).
- flexibility (thinking about how to design the space in a way that's flexible for future changes, including that these two spaces may, at some point, be combined into a single space (again, see notes)
- We'll ultimately pick a design that is creative and uses the space well. Things have changed with office space since covid, and we're trying to design a flexible space that makes the best use of the footprint possible.
- We've included some pictures that show the type of building we are hoping to build (think: farmhouse meets office). We're also looking for a design that helps us maximize economic efficiencies (ie, putting the bathrooms back to back saves costs, etc.).

Figure 58: The second related project to “Floor Plan”



Figure 59: Images to the second related project to “Floor Plan”

The third related past project with the matching rate of 85 % is shown in figure below.

Creative Floor plan for bungalow remodel / 2nd storey addition

in Architectural Design held by Bex H · Last seen: 2021-02-08 20:46:03

Open Choosing Finalist Ended

Like 0 Share 0 Tweet 0 Mark as

Post Similar Contest

Description Entries 29 Comments 32 Attachments 5

Description:
Need a simple floor plan - can be in any software or sketch by hand as long as it is to scale.

Need Main floor and 2nd floor
Looking for creative ideas to fit all the wishlist requirements
Blank floor plan with notes attached
Wishlist with photos and what I like for inspiration

Wants:
All items on wishlist included to create a 2D floor plan
Unique and creative ideas mixing modern with farmhouse/rustic
Modern open plan - but with ways of defining the zones
Good flow
Furniture will fit for the function of the space
Dimensions of spaces included

Prepaid Prize
Prize #1: **\$240**

Project Info

Created:	12 Aug 2020
Ended:	16 Aug 2020
Entries:	9
Designers:	5

[Report Contest](#)

Attachments

- Bex wishlist
- Kalowna footprint v2

Figure 60: The third related project to “Floor Plan”

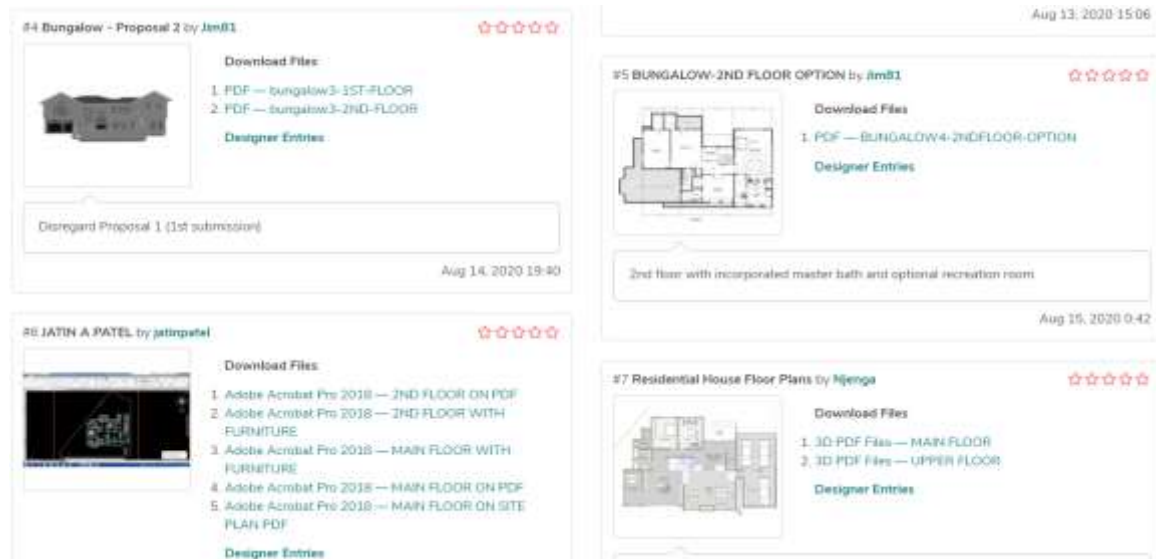


Figure 61: Images to the third related project to “Floor Plan”

7.3.3 Discussion on the result

The reason of the best matching with 91 % matching rate are the following:

- Both crowdsourcing projects are similar in the project title: Floor Plan.
- The both product requirements ask for design of large room with similar sub aspects: kitchen, living room
- Also, images from both crowdsourcing projects are having similarities in space arrangement.

The reason of the second-best matching with 87 % matching rate are the following:

- The both crowdsourcing projects are asking for at least the same main product: Floor Plan.
- The product requirements have similarities but with a different scenario: duplex on the past project
- Images from both crowdsourcing projects are having similarities.

The reason of the third best matching with 85 % matching rate are the following:

- The both crowdsourcing projects are asking for at least the same main product: Floor Plan.
- The product requirement on the past crowdsourcing project is very short, and the past project has two different floors
- Images with different floors made the image classification hard to be matched

By using all these three similar crowdsourcing pas projects as best practices / references the new user can decide the best setup as the following:

- Depend on the project budget and on the complexity of the product, the new user can decide between 240 dollars until 400 dollars.
- Short description on product requirement is powerful for motivating the participation of the crowd. It seems like that short description motivates the crowd more that longer project timeline. But it might be problematic on the end product quality, if it fulfills really the needed specification.

Chapter 8 VALIDATION OF MAIN HYPOTHESIS AND REQUIREMENTS

8.1 Requirement Satisfaction

In the first Chapter we explained that we use “Design Science” as our research methodology. We have defined the following requirements in our research methodology:

1. How to assist a manager in fetching all related text and image data effectively and easily.
2. How to prepare the data for automated analysis, including deletion of data that is irrelevant for the classification task.
3. How to develop an algorithm for the classification of text and image data.
4. The acceptance criteria: the level of accuracy for the data classification result, which should be 90%. Others include time, cost, and quality.

For the fulfillment of the first requirement, we have selected a web scraper tool and applied the necessary set up to fetch all related text and image data. The time taken for setting up the tool was 4 hours including the testing and improvisation loop. It took 5 hours to fetch all data needed from the crowdsourcing platform CADCROWD (Text and images) and 2 hours from TOPCODER (only text). In total it was 11 hours for setup and fetching process. This is in a prototyping phase of our method. Basically 11 hours is acceptable for the prototyping phase. Even for a production phase it is also acceptable. We know in which aspects we can make improvement which can make the fetching process for example 20 % faster. This can be addressed as one point for the future work.

The data fetching done during the fulfillment of the first requirement will affect the fulfillment of the second requirement. Because this was how the raw data first created. If the raw data is already clean enough on the first delivery then there will be less necessity for data preprocessing

on the second step, namely the data analysis. According to our experience we know how to optimize the fetching process, among others that some aspects on the description (requirement definition) are not to be included because they are not needed for the analysis. The tagging of the text data can also be optimized so that it complies directly the necessary data format for the data analysis step. The main data preprocessing during the second step was executed by the widget “Preprocess Text” of Orange Data Mining Tool. The existing setup works properly and we can reuse it as a best practice and also optimize it for next task. The optimization can take place on the enhancement of the “Stop word”. We can enhance it with more word elements that are not relevant for Bag of Words.

For the fulfillment of the third requirement, we have selected the best working machine learning algorithms through experiments explained on the fourth chapter. The basic setup of the algorithms gained during the manual optimization process has already fulfill the needed functionalities to classify text and images.

For the fulfillment of the fourth requirement, we have optimized the setup of Neural Network by using genetic algorithm. The overall performance was 95 %, so it fulfills the pre-defined target of 90 % classification accuracy. On the classification results the user can also get and compare information regarding time, cost, and task definition from the past crowdsourcing projects. This information is very useful for the new user so that he can adjust the step of his new crowdsourcing project based on that given best practices.

8.2 Hypothesis Evaluation

We have formulated our research hypothesis in the first Chapter as follow:

1. Best practices for performing a crowdsourced product development project are available but not commonly known by the new user / crowdsourcer
2. There are only limited methods available for handling the problem in a comprehensive way
3. Crowdsourcing project risk can be made transparent qualitatively and quantitatively
4. Artificial Intelligence can deliver a qualified performance for data mining of the complex crowdsourcing project

For the first hypothesis it is important to know and to realize the fact that crowdsourcing is still not a standard approach in the industry. The related know how is partly available in the manufacturing company and the other part of know how is mainly only available in the crowdsourcing provider. These facts are stated by Afuah [Afuah, 2017] and Nevo [Nevo, 2020]. That means for the new user (manufacturing company) the best practices are available but less than 50 %, because the main know how is to be found in the crowdsourcing provider. InnoCrowd extracted the information / best practices from the crowdsourcing provider from digital information side and made it available to the new user.

The chapter 2 addresses indirectly the second hypothesis. There are research results from the scholars explained, which are mainly focusing on the management of crowdsourcing sub aspect, such as crowd management, task management and process management. Very less researches were performed for analyzing crowdsourcing from comprehensive way. Thuan [Thuan, 2015] and Afuah [Afuah, 2017] have done this to a certain extent. InnoCrowd gives an additional contribution to the comprehensive analysis of crowdsourcing.

Chapter 7 gives a proof that crowdsourcing project risk can be made transparent qualitatively and quantitatively. It shows the result output in form of numbers for project timeline, reward and contribution number. Meanwhile the fact that there are matching past crowdsourcing projects gives a qualitative impression on the project risk of the new project. That means the crowd has experience in handling the related topic and that there were successful related past crowdsourcing projects. This information is valuable during the discussion in the manufacturing company, if they should crowdsource the project or not.

8.3 Comparison of Existing Concepts

8.3.1 Summary on Existing Concepts

8.3.1.1 Concept from Thuan

As described in the Chapter 2 there are related works which can be seen as a decision support system for crowdsourcing. But they are focused on limited aspects, namely task management and crowd management. They don't take into account other aspects such as project cost and

project timeline. By considering this limitation we compare our solution with the solution given by Thuan [Thuan, 2015] because it covers the same aspects as InnoCrowd does. Figure 62 below shows the conceptual model of the solution developed by Thuan.

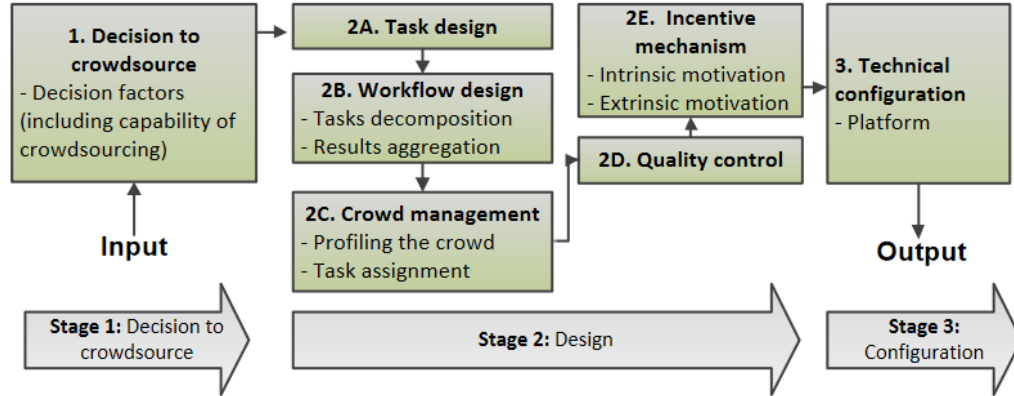


Figure 62: Conceptual model developed by Thuan

Figures below show user interface of Thuan's concept as a tool prototype. It consists of three main areas: issue, input, and advice. The decision issue is located in the left-hand side, which presents the pre-defined issues and provides an overview about the decision-making process. It allows users to navigate by a tree structure to choose an issue. The input area was presented in the (top) right-hand side. It presents a pre-defined question and parameters according to the chosen issue, as well as an explanation of the question and parameters.

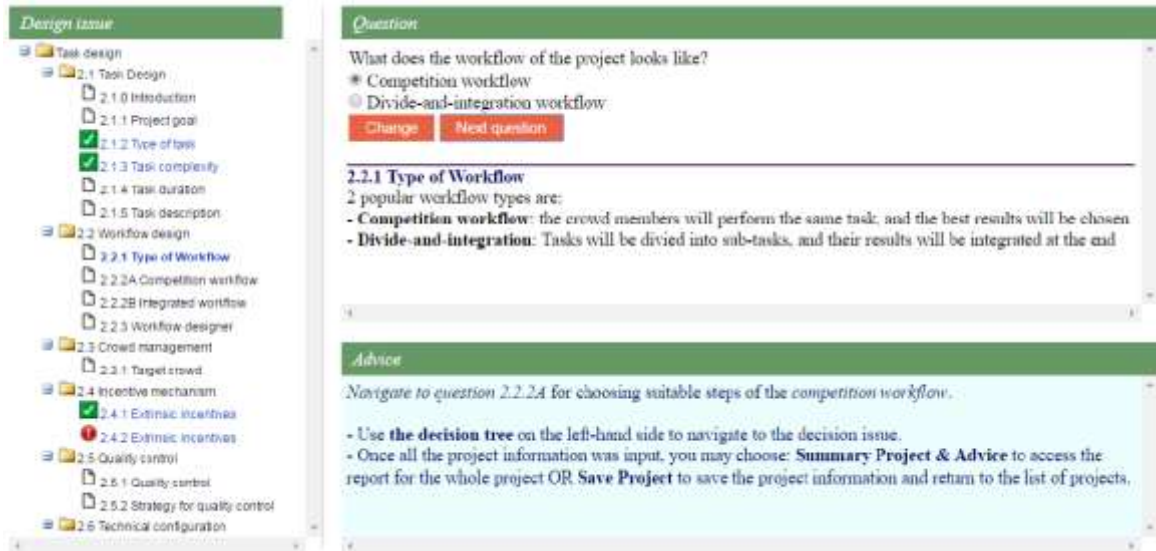


Figure 63: Design to crowdsource

The bottom right-hand side provides advice. The advice here is based on the contingent information specified by the users. Consequently, they could change when more information was provided. The complete advice was provided after the user declared all the related information.

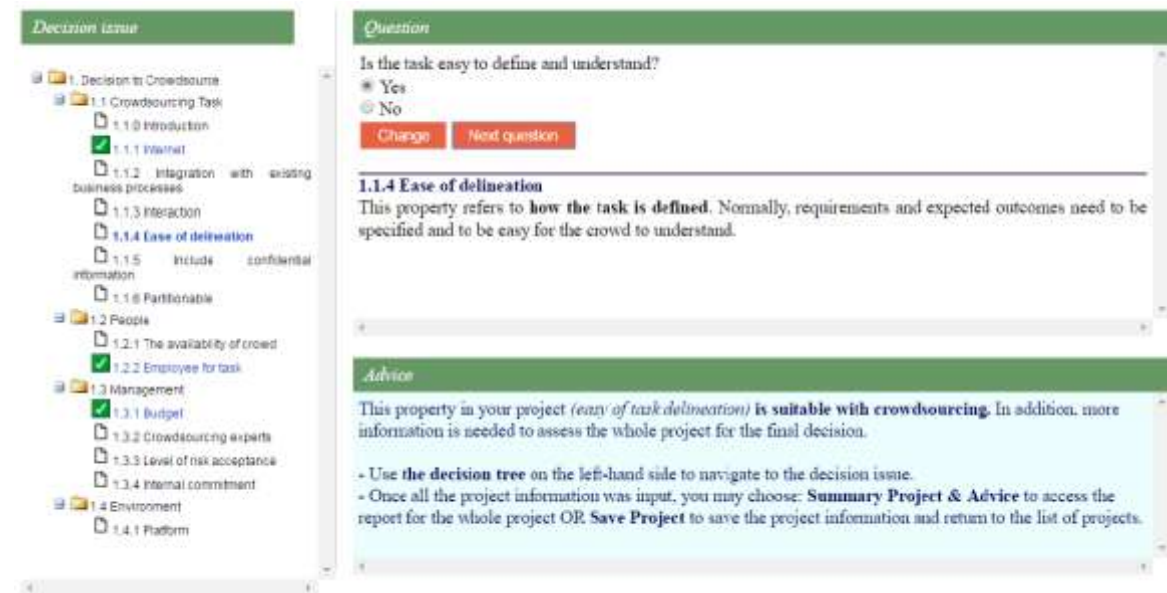


Figure 64: Process Design

The solution from Thuan is a tool to guide the manufacturing company when they want to check if they should take the crowdsourcing as their approach or not. The tool is user friendly. It gives as comprehensive suggestion to the user on the final decision. But we can see there some facts that maybe not always optimum for the new user:

- The usage is relatively time consuming because the process is not automated.
- The advices are configured by the tool at the end of the evaluation. The background information for formulating the advices can be seen as static. That means it might be not sufficient after some time because of the missing information update

8.3.1.2 Concept from Nevo

Nevo [Nevo, 2020] has developed a similar approach to the one from Thuan. He proposed that the crowdsourcing process starts with a small number of macro-level decisions that may (or may not) lead to a crowdsourcing project initiation (this step is shown as *conceptualization*).

Next, a large number of micro-level decisions are triggered that focus on project *initiation* and *execution*. The final step evaluates the impact of crowdsourcing and its measurable value to the firm in the short and long term (*termination* and *ongoing* steps). Each step in this process requires the organization to make choices, many of which relate to the strategic goals of the organization in question.

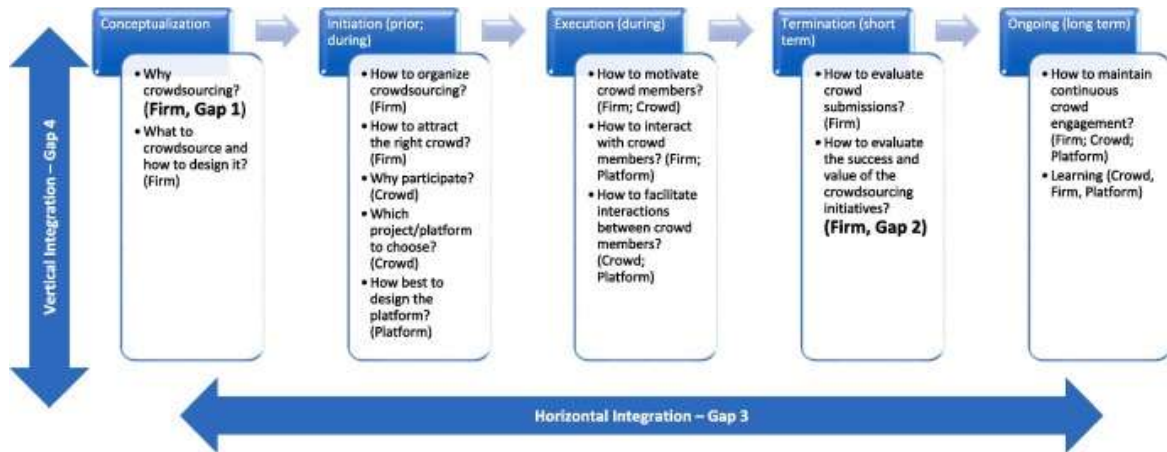


Figure 65: A Road Map for Crowdsourcing in the Organizational Context

His work is based on a comprehensive study of related research activities for crowdsourcing. If we try to make a grouping to that 5 crowdsourcing evaluation phases we can take the first and the fourth phases as phases specific for the new user (internal the manufacturing company). Meanwhile the second, third and fifth phases are relevant phases for the crowdsourcing provider.

The road map is a good start for the company to assess the risk of crowdsourcing. Because the company will know the complexity of the crowdsourcing process and he also get informed on some critical questions so that he can be more prepared with the risk behind.

InnoCrowd gives a qualitative and quantitative answer to main points from the phase one and phase 4 of the road map.

8.3.1.3 Concept of InnoCrowd

For a comparison in the high level, we show InnoCrowd's conceptual model in the figure 66 below. InnoCrowd and Tuan's solution are to be used as tool for a decision support in crowdsourcing project.

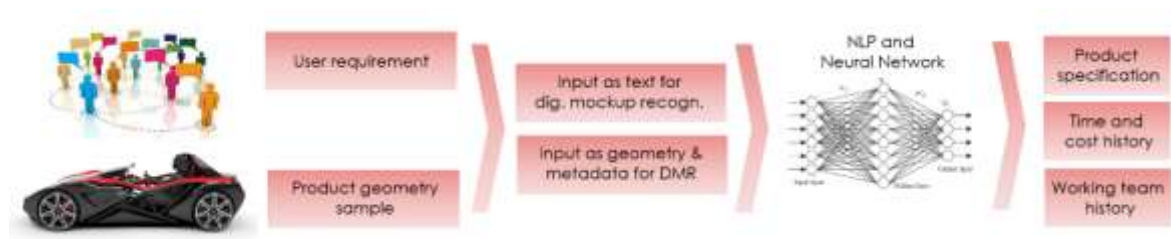


Figure 66: InnoCrowd's conceptual model

In practice both solutions deal with the aspects task / product specification, cost management, as well crowd management. Solution of Thuan takes quality control into account. Meanwhile InnoCrowd seals with time management.

8.3.2 Comparison Method

We need to compare the existing solution by using a standardized method. The following sub chapters describes the methods available for comparing IT solutions.

8.3.2.1 ISO / IEC 9126

ISO / IEC 9126 is an international standard that specifies criteria for the quality and evaluation of software products.

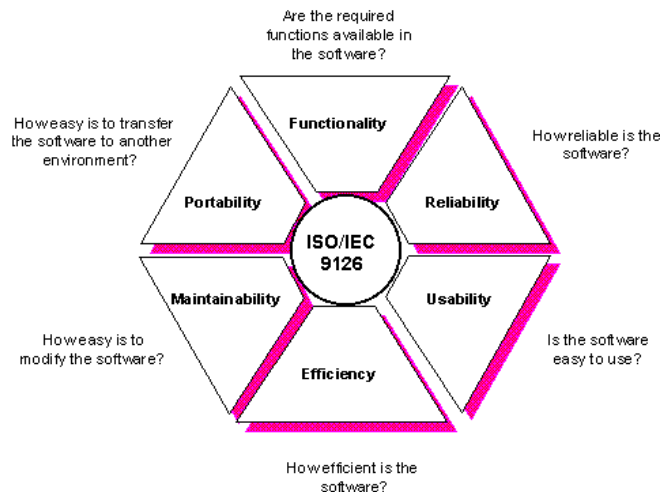


Figure 67: ISO / IEC 9126

It recommends to pay attention to the following aspects when evaluating a software:

Maintainability: How much effort does the implementation of specified changes to the software require? - The effort required to carry out specified changes. Changes can include corrections, improvements or adjustments to changes in the environment, requirements or functional specifications.

Portability: How easily can the software be ported to another environment? - Suitability of the software so that it can be transferred from one environment to another. The environment can be an organizational, hardware or software environment.

Functionality: To what extent does the software have the required functions? - Presence of functions with defined properties. These functions meet the defined requirements.

Reliability: Can the software maintain a certain level of performance under certain conditions over a certain period of time? - Ability of the software to maintain its level of performance under specified conditions for a specified period of time.

Usability: How much effort does the use of the software require of the users and how is it assessed by them? - Effort required for use and individual assessment of use by a specified or required user group. See also software ergonomics.

Efficiency: What is the relationship between the performance level of the software and the resources used? - Relationship between the level of performance of the software and the scope of the operating resources used under specified conditions.

8.3.2.2 Boehm's Model

Boehm proposed a model of software quality factors, where the quality factors are represented in a hierarchical manner [Boehm, 1976]. The high-level factors are associated with the software use, while the low-level factors are related to metrics. The model is shown in Figure below. The group of factors associated to As-is Utility determines that the software be reliable, efficient and has some usability aspects (human-engineered). According to Boehm and others, As-is-Utility, Maintainability and Portability are necessary, but not sufficient conditions to General Utility. This is because some domain specific softwares require additional factors not considered by the model (e.g. security).

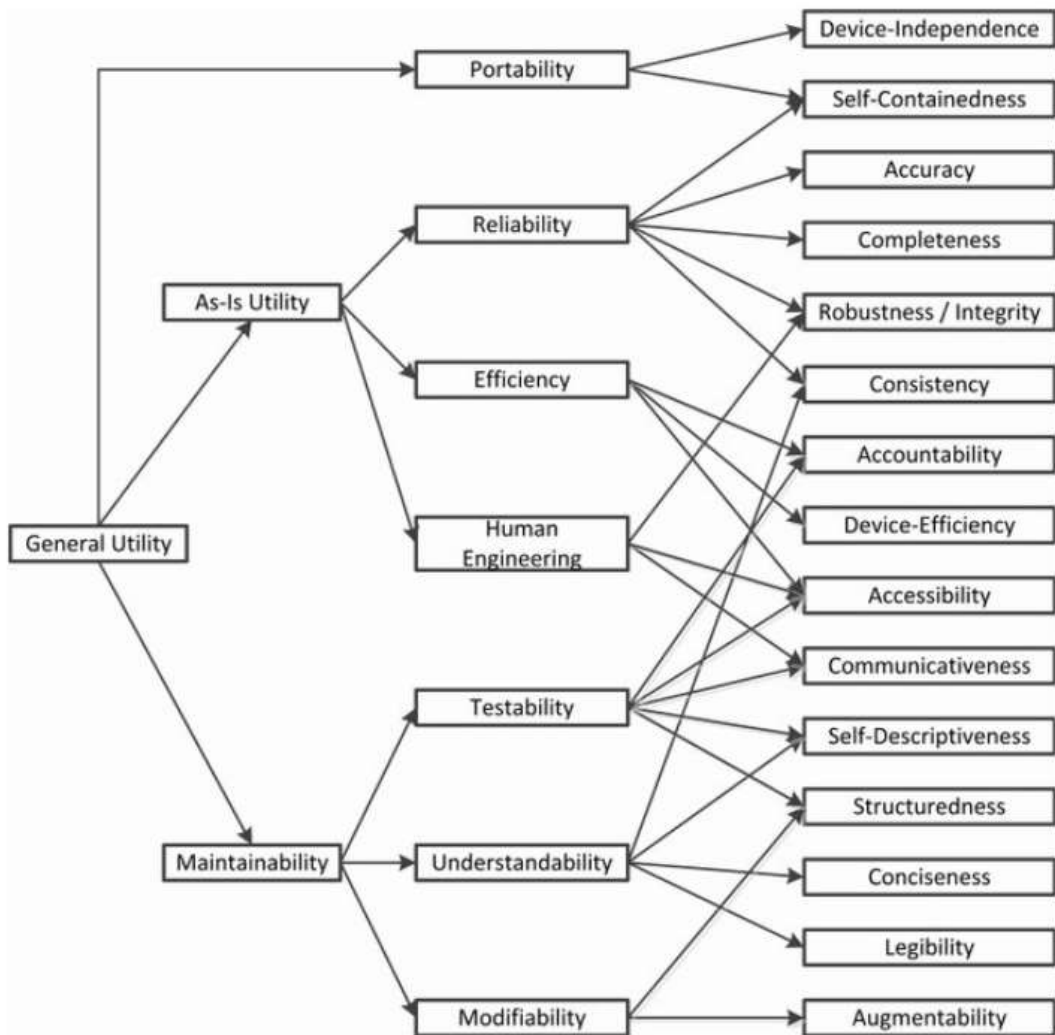


Figure 68: Boehm's model

8.3.2.3 Similarity Measurement

Many approaches have been proposed in different domains in order to measure the similarity: in the engineering of information systems within the component re-use, in the software development to ensure the traceability of the code, in the framework of document analysis to index and retrieve documents, or for the schema analysis of heterogeneous database [Rohleder, 2009]. We will use part of the similarity measurement method for comparing our solution with other available solutions.

In the context of similarity measurement, it is useful to classify products into product families that exploit common characteristics of related products. Most of the product can be classified according to two dimensions of scope, i.e. quality scope and domain scope. The quality scope (profile) captures almost all common and different characteristics of the product family members (configurable product family of a software). Figure 69 shows the Quality Profile of Final Product for ISIC Entertainment and Product Lifecycle Management (PLM) Product Report.

Quality Profile of Final Product for a Report Tool for ISIC classification Entertainment

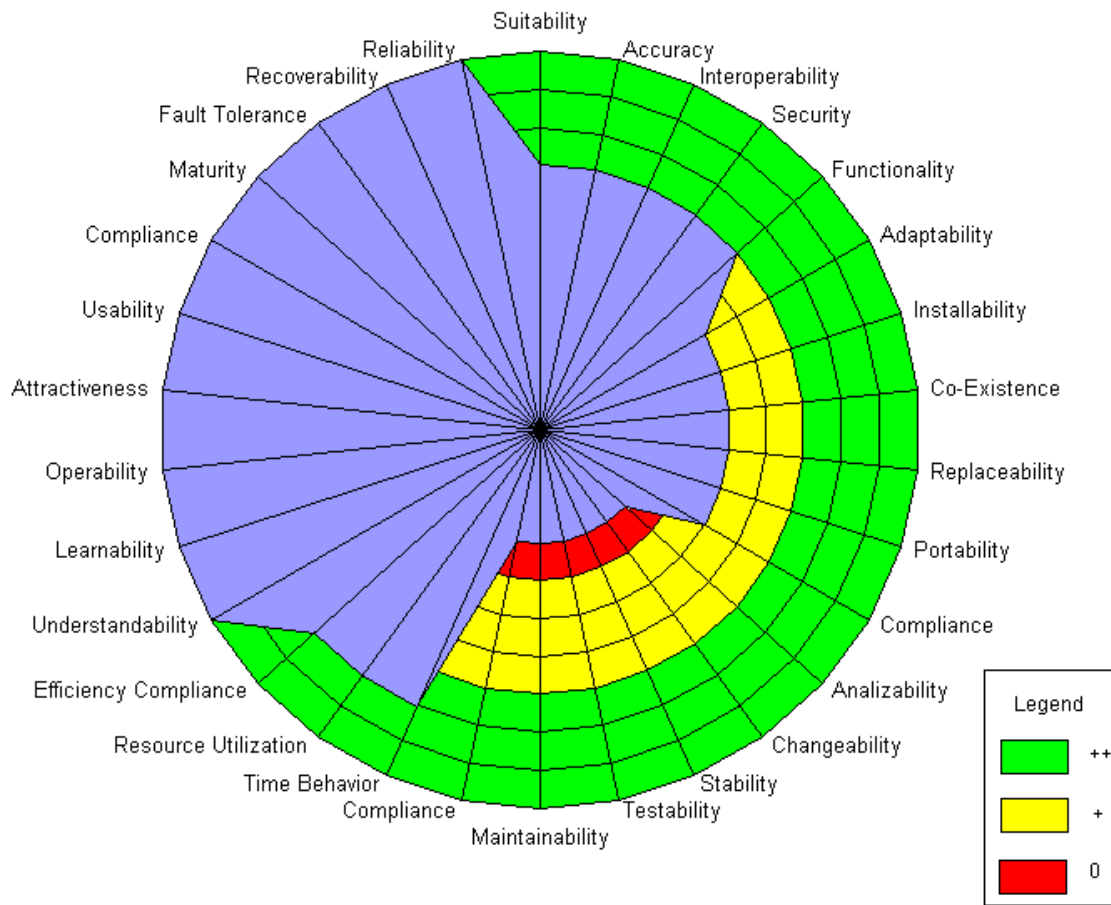


Figure 69: Quality Profile of Final Product for ISIC Entertainment and Report Tool Product

8.3.3 Applying the comparison method

We compared the existing methods explained in the previous sub chapters. We have found many similarities existing in those three methods in the following aspects:

- Maintainability
- Portability
- Functionality
- Reliability

- Usability
- Efficiency

These six aspects are covered completely in the ISO / IEC 9126. Therefore, we think that IEC 9126 is a representative for all three methods described in the previous sub chapter. We use this method to compare InnoCrowd with other existing solutions.

Thuan has provided a tool as a prototype for his concept. Nevo has provided only a guideline (road map) for the evaluation of the crowdsourcing approach and not IT tool is provided by him. Therefore, it not possible to compare InnoCrowd one to one with the solution from Nevo. We compare InnoCrowd with the solution from Thuan by using ISO / IEC 9126 which can be broken down in details as shown in the figure below.

Characteristics	Subcharacteristes
Functionality	suitability
	accuracy
	interoperability
	security
	functionality compliance
Reliability	maturity
	fault tolerance
	recoverability
	reliability compliance
Usability	understandability
	learnability
	operability
	attractiveness
	usability compliance
Efficiency	time behaviour
	resource utilisation
	efficiency compliance
Maintainability	analysability
	changeability
	stability
	testability
	maintainability compliance
Portability	adaptability
	installability
	co-existence
	replaceability
	portability compliance

Table 2: ISO / IEC 9126 in detail

Maintenanceability

- **Analyzability:** The effort required to diagnose defects or causes of failure or to determine parts that need to be changed.

- **Maintainability compliance:** The degree to which the software complies with standards or agreements on changeability.
- **Changeability:** The effort involved in making improvements, eliminating errors, or adapting to changes in the environment.
- **Stability:** the likelihood of unexpected effects of changes occurring.
- **Testability:** the effort required to test the changed software.

Usability

- **Attractiveness:** The attraction of the application to the user.
- **Usability:** effort for the user to operate the application.
- **Learnability:** effort for the user to learn the application (e.g. operation, input, output).
- **Operability:** The degree to which the software meets standards or usability agreements.
- **Understandability:** effort for the user to understand the concept and the application.

Efficiency

- **Conformity:** The degree to which the software meets standards or agreements on efficiency.
- **Efficiency compliance:** response and processing times as well as throughput during function execution.
- **Resource utilization:** number and duration of the resources required to fulfill the functions. Resource consumption, such as CPU time, hard disk access, etc.

Functionality

- **Suitability:** Suitability of functions for specified tasks, for example task-oriented composition of functions from sub-functions.

- **Security:** Ability to prevent unauthorized access, both accidental and deliberate, to programs and data.
- **Interoperability:** Ability to interact with given systems.
- **Conformity:** Ability of the software product to comply with standards, conventions or legal provisions and similar regulations related to functionality.
- **Compliance:** Characteristics of software that cause the software to comply with application-specific standards or agreements or legal provisions and similar regulations.
- **Correctness:** Delivering the correct or agreed results or effects, for example the required accuracy of calculated values.

Portability

- **Adaptability:** Ability of the software to adapt it to different environments.
- **Interchangeability:** Possibility of using this software instead of a specified other in the environment of that software, as well as the effort required for this.
- **Installability:** The effort required to install the software in a specified environment.
- **Coexistence:** Ability of the software to work alongside another with similar or identical functions.
- **Conformity:** The degree to which the software meets standards or agreements on portability.

Reliability

- **Fault tolerance:** the ability to maintain a specified level of performance in the event of software errors or non-compliance with its specified interface.
- **Conformity:** The degree to which the software meets standards or agreements on reliability.
- **Maturity:** Low frequency of failure due to fault conditions.

- **Recoverability:** the ability, in the event of a failure, to restore the level of performance and recover the data directly affected. The time and effort required for this must be taken into account.

Subcharacteristics	InnoCrowd vs Thuan's solution	Reason
Suitability	Neutral	-
Accuracy	InnoCrowd is better	Based on real life data with numbers
Interoperability	Neutral	-
Security	Neutral	-
Funtionality compliance	Neutral	-
Maturity	InnoCrowd is better	Based on real life data with numbers
Fault Tolerance	InnoCrowd is better	Based on real life data with numbers
Recoverability	Thuan' solution is better	Minimum IT process in background
Reliability compliance	Neutral	-
Understandability	Neutral	-
Learnability	Neutral	-
Operability	Neutral	-
Attractiveness	InnoCrowd is better	Automatic data processing
Usability compliance	Neutral	-
Time behaviour	InnoCrowd is better	Automatic data processing
Resource utilisation	Neutral	-
Efficiency compliance	InnoCrowd is better	Automatic data processing
Analysability	Thuan' solution is better	Using more human intelligence
Changeability	Thuan' solution is better	Using more human intelligence
Stability	Neutral	-
Testability	Neutral	-
Maintainability compliance	Thuan' solution is better	Minimum IT process in background
Adaptability	Neutral	-
Installability	Neutral	-
Coexistence	Neutral	-
Replaceability	Neutral	-
Portability compliance	Neutral	-

Table 3: Comparison between InnoCrowd and Thuan's solution

Table 4 shows a qualitative comparison between InnoCrowd and Thuan's solution. The value "Neutral" means that the both solutions are having approximately the same performance. When the value is not neutral then there is an explanation on the column "Reason". The comparison might be subjective but at least the given reasons are relative logic and acceptable. As the result of the comparison we can see that InnoCrowd has 6 times better points,

meanwhile Thuan's solution get 4-time better points. That means qualitatively InnoCrowd is better than Thuan's solution

8.4 Threats to Validity

We can consider the following points as threats to validity:

- The deep learning neural network for the image analysis was trained before with a very large and standard image dataset (by Orange Data Mining, the tool provider) and then it is enhanced with the training on the image dataset from the crowdsourcing platform. When InnoCrowd is to be used to analyze data from other crowdsourcing platform then the result validity of image analysis might be lower. Therefore, a certain training session should be performed before to the dataset from the actual crowdsourcing platform in addition.
- The same case also for the text analysis. The usage for other crowdsourcing platform will need a retraining of the NLP module of InnoCrowd.

8.5 Open Issues

We can consider the following point as open issue:

- During the dataset generation we use a web scrapper tool in order to fetch all related data from the website of the crowdsourcing provider. We need to take into account that the usage of a web scrapper tool for data collection is not always allowed. To a certain circumstance it might be against the law. Therefore, it is to be checked thoroughly if this procedure is to be used to other websites.

Chapter 9 RESULTS AND CONCLUSIONS

9.1 Summary

Our lives are increasingly influenced by digital technologies. Digitalization is changing both our society and us as individuals - and including the way we work and how companies organize their work and service creation processes. With crowdsourcing and crowd work is an innovative model for organizing product innovation creation process. It can be assumed that more and more companies are adopting the concept in order to be able to access a large pool of labor quickly and in a targeted manner. In the medium to long term, this will result in a disruptive change in organizational and work structures.

However, the change is not always as fast as expected, especially when many stakeholders are involved and also when the process to execute a crowdsourced product development is relatively complex. This phenomenon is proven by research results discussed in this work. One of the tasks of this work is to identify problems and challenges related to the application of crowdsourcing in the standard work of the organization (manufacturing company). And then to identify what has been done so far by the scholars to encounter the problem.

This PhD work tried to contribute to give a solution by developing a new method called InnoCrowd. It is an artificial intelligence based decision support system for a crowdsourced product development. InnoCrowd should be able to help organization with the new crowdsourcing project initiation. The organization, who is not familiar or doesn't know best practices of Crowdsourcing will get a relatively quick access to information on the related past project, so that the project risk is better known by taking into account experiences of other crowdsourcer in the past.

InnoCrowd has a knowledge-based classification system built up with machine learning and Natural Language Processing algorithms. Dataset from the real-life data from crowdsourcing platform and train the machine learning algorithm with this The dataset is generated in order

to build up the classification system. We verify the solution concept by using dataset from crowdsourcing platform CAD Crowd and TopCoder. The verification result on product requirement analysis (as free text file) gives a good classification result, namely in the accuracy level between 90% and 91%. Meanwhile the verification result on digital model analysis gives even a higher accuracy, between 95% and 98%. The overall performance when the text and image analysis are integrated is 90%.

In order to optimize further the algorithm and also to avoid overfitting we used genetic algorithm for enhancing the parameter of the machine learning. After the optimization with genetic algorithm InnoCrowd can perform an accuracy level 95%. The comparison of InnoCrowd with other solution gives a clear result, that InnoCrowd performs better.

In relation with the usage of InnoCrowd as a decision support system then it is technically done by providing the user the information related to the previous project matched with the planned project. Three industrial use cases by using real life data are shown. The result from InnoCrowd analysis is given. The user will have access to the information of the previous project cost, the time needed to finish the project and also the crowd workers who had work for similar task. These information give a clear transparency on the project risk, so that InnoCrowd can be used as an answer to the management when they want to crowdsource a product development project.

9.2 Future Work

So far, we have developed the basic functionality of InnoCrowd. What we still have to develop in the future is the user usability so that the process chain starting from data fetching, data preprocessing, data classification and result presentation can be done in an integrated environment. And we still need to optimize the processing time as well as the accuracy of the algorithm.

We collected all necessary improvement from each step, among others the preprocessing during the data generation and the preprocessing during the data mining. The realization of this improvement will be in form of windows script because no tool is available for handling this requirement directly. The requirement is related with the modification of the raw data into the compatible format of the next data processing step.

One aspect that is still not handled thoroughly in the current InnoCrowd prototype is the result presentation of the matching past crowdsourcing projects. In order to handle this, we need to build up a database for storing each parameter of the past crowdsourcing project.

In order to make InnoCrowd become a generic solution we also need to apply dataset from other crowdsourcing providers. We train InnoCrowd with their data and we can test scenarios for other domain, for example in the domain of electrical engineering and even health care.

REFERENCES

[**Achi, 2016**] Abdelkader Achi, Camille Salinesi, Gianluigi Viscusi: Information Systems for Innovation: A Comparative Analysis of Maturity Models' Characteristics. pp. 78-90. CAiSE Workshops 2016

[**Achtaich, 2018**] Achtaich, A., Souissi, N., Mazo, R., Roudies, O., Salinesi, C.: A DSPL Design Framework for SASS: A Smart Building Example. EAI Endorsed Trans. Smart Cities, vol. 2, no. 8, p. 154829 (2018).

[**Afuah, 2017**] Afuah, A.: Co-opetition in crowdsourcing: When simultaneous cooperation and competition deliver superior solutions. In C. Tucci, A. Afuah, & G. Viscusi (eds.), *Creating and Capturing Value Through Crowdsourcing*, pp. xx–yy. Oxford: Oxford University Press (2017)

[**Ali, 2019**] Amir Ali: *An Intuitive Guide of Artificial Neural Network with Practical Implementation in Keras & Tensorflow Artificial Neural Network* (2019)

[**Allahbakhsh, 2013**] Mohammad Allahbakhsh: Quality Control in Crowdsourcing Systems: Issues and Directions. IN: *IEEE Internet Computing* 17(2):76-81 (2013)

[**Allen, 1976**] Allen, R., Fowler, F.: *The pocket Oxford dictionary of current English*. Oxford University Press (1976).

[**Andre', 2012**] André, S.: CrowdWeaver: Visually Managing Complex Crowd Work. In: *CSCW '12: Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (2012)

[Appaswamy, 2018] Niranjan Appaswamy: Ensemble of Decision Stump, Decision Tree and Linear Regression for Efficient Fraud Detection using Stacking, In: Conference: ICCII (2018)

[Bakir, 2001] Saad Bakir: Classification of Distribution-free quality controls charts. In: Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, (2001)

[Berrar, 2018] Daniel Berrar: Bayes' Theorem and Naive Bayes Classifier. In book: Reference Module in Life Sciences (2018)

[Bezdan, 2019] Timea Bezdan: Convolutional Neural Network Layers and Architectures. In Conference: Sinteza (2019)

[Boehm, 1976] Boehm, B.W., Brown, J.R., Lipow, M.: Quantitative evaluation of software quality. In: 2nd International Conference on Software Engineering. pp. 592–605. IEEE, Los Alamitos (1976)

[Borchert, 2012] Borchert, K., Hirth, M., Schnitzer, S., Rensing, C.: Impact of Task Recommendation Systems in Crowdsourcing Platforms. In: Proceedings of the First International Workshop on Crowdsourcing and Data Mining (2012).

[Buhse, 2011] Buhse, W., Reppesgard, L., Lessmann, U.: Der Case Local Motors: Co-Creation and Collaboration in der Automotive-Industrie. (2011).

[Chang, 2018] Ming-Wei Chang: Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. IN AI-Googleblog (2018)

[Chatterjee, 2014] Chatterjee, S., Khandekar, P., Kumar, B.: Reimagining enterprise innovation through crowdsourcing. (2014).

[Chengsheng, 2017] Tu Chengsheng: AdaBoost typical Algorithm and its application research. In: MATEC Web of Conferences 139(2):00222 Follow journal (2017)

[Cheung, 2012] S. Cheung, "How Companies Can Leverage Crowdsourcing", Working Paper CISL#2012-2, Cambridge, February 2012

[Ciocca, 2006] Gianluigi Ciocca: Supervised And Unsupervised Classification Post-Processing for Visual Video Summaries. In: IEEE Transactions on Consumer Electronics 52(2):630 - 638, IEEE Xplore (2006)

[Cui, 2020] Lixin Cui: The study of customer involved service innovation under the crowdsourcing : a case study of MyStarbucksIdea.com. In: Journal of industry - university collaboration : JIUC. - Bingley : Emerald, ISSN 2631-357X, ZDB-ID 3001421-9. (2020)

[Cutler, 2011] Adele Cutler. In: Random Forests. In book: Ensemble Machine Learning: Methods and ApplicationsChapter. Publisher: Springer (2011)

[Ding, 2013] Shifei Ding: Evolutionary artificial neural networks: A review. In: Artificial Intelligence Review 39(3) Follow journal (2013)

[Eagle, 2009] Eagle, N.: txteagle: Mobile Crowdsourcing. pp. 447–456.Springer, Berlin, Heidelberg (2009).

[Ebert, 2014] Christof Ebert: Collaborative Software Development Platforms for Crowdsourcing. IN: IEEE Software 31(2):30-36 (2014)

[fcagroup,2012]

https://www.fcagroup.com/investorday/PresentationList/Global_Architectures_and_Standardization.pdf

[Figuerola, 2001] Carlos G. Figuerola: Automatic vs. Manual categorisation of documents in Spanish. In: Journal of Documentation (2001)

[Ford, 2015] Robert C Ford: Crowdsourcing: A new way of employing non-employees? . In: Business Horizons 1(5):377-388 (2015)

[Gassmann, 2012] O. Gassmann, "Crowdsourcing - Innovationsmanagement mit Schwarmintelligenz - Interaktiv Ideen finden - Kollektives Wissen effektiv nutzen – Mit Fallbeispielen und Checklisten", Carl Hanser Verlag, München, (2012)

[Gartner, 2014] Gartner, Inc.: Predicts 2015: Sourcing Strategies Shift From ‘Built to Last’ to ‘Built to Adapt’ (2014)

[Geiger, 2012] Geiger, D.: Crowdsourcing information systems - definition, typology, and design. In: Conference: ICISAt: Orlando (2012)

[Gibb, 2015] Alicia Gibb: Building Open Source Hardware: DIY Manufacturing for Hackers and Makers. Addison-Wesley (2015)

[Heinert, 2010] Michael Heinert: Support Vector Machines – Teil 1: Ein theoretischer Überblick. In: Zeitschrift für Geodäsie, Geoinformation und Landmanagement 135(3):179–189 (2010)

[Hirth, 2012] Matthias Hirth: Analyzing Costs and Accuracy of Validation Mechanisms for Crowdsourcing Platforms. In Mathematical and Computer Modelling (2012)

[Howe, 2008] Howe, J.: Crowdsourcing: How the power of the crowd is driving the future of business. (2008).

[Kittur, 2011] A. Kittur, B. Smus, S. Khamkar, R.E. Kraut, "CrowdForge: Crowdsourcing complex work", Proceedings of the ACM Symposium on User Interface Software and Technology. New York: ACM Press.UIST 2011)

[Kumar, 2012] Yugal Kumar: Analysis of Parametric & Non Parametric Classifiers for Classification Technique using WEKA. In: International Journal of Information Technology and Computer Science 4(7):43-49 (2012)

[Kusumah, 2018] Indra Kusumah, Camille Salinesi, Clotilde Rohleder: Platform for crowdsourcing in automobile industry. In: 24th ICE/IEEE International Technology Management, Stuttgart (2018)

[Kusumah, 2020] Indra Kusumah, Camille Salinesi, Clotilde Rohleder: InooCrowd, a contribution to anIoT based Engineering Product Development. In: 9th International Conference on Computer Science and Information Technology (CoSIT 2020), Zurich (2020)

[Kusumah, 2021] Indra Kusumah, Camille Salinesi, Clotilde Rohleder: InnoCrowd, an AI based optimization of a crowdsourced product development. In: IFIP 18th International Conference on Product Lifecycle Management in Brazil (2021)

[Knop, 2019] Nicolas Knop: Internes Crowdsourcing – Herausforderungen und Lösungsstrategien für eine erfolgreiche Transformation der Arbeitsorganisation. IN: HMD Praxis der Wirtschaftsinformatik Follow journal (2019)

[Li, 2017] Yang Li: Word Embedding for Understanding Natural Language: A Survey. In book: Guide to Big Data ApplicationsEdition. Publisher: Springer International (2017)

[Li, 2020] Ziyuan Li: CrowdSFL: A Secure Crowd Computing Framework Based on Blockchain and Federated Learning. In. Electornics 9(5): 773 (2020)

[Lin, 2017] Fan Lin: An artificial neural network approach for screening test escapes. In: 22nd Asia and South Pacific Design Automation Conference (ASP-DAC). (2017)

[Little, 2009] Greg Little: TurKit: Tools for Iterative Tasks on Mechanical Turk. In: Proceedings of the ACM SIGKDD Workshop on Human Computation, Paris, France, (2009)

[Local Motors, 2004] Local Motors Homepage. <https://localmotors.com/>.

[Malone, 2011] Malone, T.: The Big Idea: The Age of Hyperspecialization.
In: Harvard Business Review (2011)

[Martin, 2008] Martin, N.: Crowdsourcing: Systematisierung praktischer Ausprägungen und verwandter Konzepte. In: Tagungsband der Multi-Konferenz Wirtschaftsinformatik (2008)

[Mongiello, 2016] Mongiello, M., Boggia, G., Di Sciascio, E.: ReIOS: Reflective Architecting in the Internet of Objects. Proc. 4th Int. Conf. Model. Eng. Softw. Dev., pp. 384–389 (2016).

[Nevo, 2020] Dorit Nevo: Crowdsourcing as a strategic IS sourcing phenomenon: Critical review and insights for future research. In: The Journal of Strategic Information Systems (2020)

[Peppers, 2012] Peppers: Design Science Research in Information Systems: Advances in Theory and Practice. In: Proceeding of 7th International Conference, DESRIST 2012, Las Vegas, NV, USA, May 14-15, (2012)

[PLM 2014] <http://www.product-lifecycle-management.info/what-is-plm/The-PLM-Value-Framework.pdf>

[plm Siemens, 2011] https://www.plm.automation.siemens.com/de_de/Images/Siemens-PLM-Teamcenter-Mobility-fs_tcm73-121616.pdf

[Pohl, 2005] Pohl, K., Böckle, G., Van Der Linden, F.: Software Product Line Engineering. Foundations, Principles, and Techniques, vol. 49, no. 12. (2005).

[Prpic, 2015] Prpic, J., Shukla, P., Kietzmann, J., McCarthy, I.: How to Work a Crowd: Developing Crowd Capital through Crowdsourcing. In: Business Horizons, pp 77-85 (2015).

[Qudar, 2020] Mohiuddin Qudar: A Survey on Language Models. In: datalab.science (2020)

[Renade, 2015] Renade, G. Varshney, L.: To Crowdsourc or not to Crowdsourc? In: HCOMP@AAAI (2012)

[Rohleder, 2009] Rohleder C.: Quality Product Derivation: A Case Study for Quality Control at Siemens, Best Paper Award of 8th WSEAS (World Scientific and Engineering Academy and Society) International Conference on Software Engineering, Parallel and Distributed Systems (SEPADS '09), February 2009, Cambridge, UK, pp. 51-56 (2009)

[Roth, 2014] Yannig Roth: Crowdsourcing In The FMCG Sector - Development And Overview. In: ISPO's "Open Innovation Wiesn" (2014)

[Rucco, 2018] Matteo Rucco: A methodology for part classification with supervised machine learning. In: Artificial Intelligence for Engineering Design Analysis and Manufacturing 33(1):1-14 (2018)

[Sautter, 2011] Guido Sautter: High-Throughput Crowdsourcing Mechanisms for Complex Tasks. In: SocInfo (2011)

[Schmitz, 2016] Heinz Schmitz: It's about time: Online Macrotask Sequencing in Expert Crowdsourcing, Trier University of Applied Sciences (2016)

[Salinesi, 2016] Camille Salinesi: Reusable knowledge in security requirements engineering: a systematic mapping study. In: Requirements Engineering 21 (2), 251-283 (2016)

[Sarvepalli, 2015] Sarat Kumar Sarvepalli: Deep Learning in Neural Networks: The science behind an Artificial Brain. In: Liverpool Hope University (2015)

[Sawalhi, 2014] Nabil El Sawalhi: A Neural Network Model for Building Construction Projects Cost Estimating. In: Journal of Construction Engineering and Project Management 4(2014)

[Schade, 2019] Frauke Schade: Age of the Customer: Interne und externe Daten zur Beschreibung von Zielkunden und ihre Eignung für die Marktsegmentierung. In Book: Praxishandbuch Informationsmarketing (2019)

[Schenk, 2009] Schenk, E., Guittard, C.: Crowdsourcing: What can be Outsourced to the Crowd, and Why? (2009).

[Schenk, 2012] Sawyer, P., Mazo, R., Diaz, D., Salinesi, C., Hughes, D.: Constraint Programming as a Means to Manage Configurations in Self-Adaptive Systems. Spec. Issue IEEE Comput. J. Dynamic Softw. Prod. Lines. vol. 45, pp. 56–63 (2012).

[Shilovitsky, 2012] Oleg Shilovitsky: The future of PLM business models. In: BeyondPLM (2012)

[Sperandei, 2014] Sandro Sperandei: Understanding logistic regression analysis. In: Biochemia Medica 24(1):12-8 (2014)

[Su, 2015] H. Su, S. Maji, E. Kalogerakis, E.L. Miller, "Multi-view Convolutional Neural Networks for 3D Shape Recognition", Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), 2015, Washington, pp 945-953

[Thanh, 2015] Tran-Thanh: Crowdsourcing complex workflows under budget constraints. In: AAAI Conference on Artificial Intelligence (2015)

[Thuan, 2015] Thuan, N., Antunes, P., Johnstone, D., Xuan, H., Hoang, N.: Building an Enterprise Ontology of Business Process Crowdsourcing: A Design Science Approach. In PACIS 2015 Proceedings, vol 112 (2015).

[TopCoder, 2015] <https://www.topcoder.com/>

[Tripathi, 2011] Tripathi, K.P.: Decision Support System Is a Tool for Making Better Decisions in the Organization. In: Indian Journal of Computer Science and Engineering 2 (2011)

[Urban Hub, 2015] <http://www.urban-hub.com/ideas/say-goodbye-to-downtime-and-delays-with-predictive-smart-technologies> (2015)

[Vaisch, 2017] Rajan Vaish: Twitch crowdsourcing: Crowd contributions in short bursts of time. In ACM CSCW 17, Portland, USA (2017)

[Valantin, 2018] Lima Valantin: Why you should not trust only in accuracy to measure machine learning performance. In: Medium.com [2018]

[Valentine 2014] Valentine, M.: Expert crowdsourcing with flash teams. In: Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology, UIST '14, pages 75–85, New York, NY, USA, (2014)

[Verspoor, 2013] Karin Maria Verspoor: Natural Language Processing. In book: Encyclopedia of Systems Biology (2013)

[Wang, 2019] Qiang Wang: Learning Deep Transformer Models for Machine Translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (2019)

[Weforum, 2021] <https://reports.weforum.org/digital-transformation/delivering-change-digital-transformation-in-logistics>

[Whittle, 2010] Whittle, J., Sawyer, P., Bencomo, N., Cheng, B., Bruel, J.: RELAX: a language to address uncertainty in self-adaptive systems requirement. Requirement Engineering, vol. 15, no. 2, pp. 177–196, Jun. (2010).

[Wikipedia, 2014] https://en.wikipedia.org/wiki/Ubiquitous_computing

[Yang, 2016] Yang, Y., Karim, M., Saremi, R., Ruhe, G.: Who Should Take This Task? – Dynamic Decision Support for Crowd Workers. In: Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, pp. 1–10 (2016).

[Yu, 2019] Dunhui Yu: Crowdsourcing Software Task Assignment Method for Collaborative Development. In: IEEE Access PP(99):1-1 (2019]

[Yuen, 2011] Man-Ching Yuen: A Survey of Crowdsourcing Systems. In Conference: PASSAT/SocialCom 2011, Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom), Boston, (2011)

[Zhang, 2010] Yin Zhang: Understanding bag-of-words model: a statistical framework: In: International Journal of Machine Learning and Cybernetics (2010)

[Zhao, 2015] Zhao, Z., Wei, f., Zhou, M., Chen, W., Ng, W.: Crowd-Selection Query Processing in Crowdsourcing Databases: A Task-Driven Approach. In: EDBT Computer Science (2015).

[Zheng, 2016] Zheng, Y., Li, G., Cheng, R.: DOCS: DomainAware Crowdsourcing System. In: Proceedings of the VLDB Endowment, pp 361-372 (2016).