AI-Based Requirement Elicitation for Product Innovation in Data-Driven Marketing

A Case Study

[1.]Clotilde Rohleder, [2.]Indra Kusumah, [3.]Camille Salinesi, [4.]Michael Meier

[1,2,]*University of Applied Sciences ConstanceConstance, Germany*
[3.]University of Paris 1 – Panthéon – SorbonneParis, France
[4.]Schindler / Parent GmbHMeersburg, Germny

ABSTRACT

The rapid digitalization of the economy has significantly bolstered digital businesses, especially in the wake of the Covid-19 pandemic. An essential component of this growth is innovation. As e-commerce users frequently share their experiences and aspirations on social media, there is a treasure trove of data that can be harnessed for product innovation. While this customer voice data is invaluable, the sheer volume makes manual analysis impractical. The challenge lies in efficiently extracting and analyzing this data to facilitate early innovation and improvements in e-commerce platforms. This paper introduces an AI-based semi-automated method for requirements elicitation, employing machine learning algorithms to analyze data from user reviews in the e-commerce and social media. Our approach aims to identify potential product or service innovations by using data and machine learning. This approach combines different domains like marketing, product innovation, requirements engineering (RE) and machine learning. It is an artificial intelligence (AI)-based method in the marketing field which generates requirements knowledge gained from existing data in order to promote innovation of user/customer adopted products. Innovation has been defined as the creation and delivery of novel (or renewed) processes, products, services that result in significant outcomes reflected by user/customer adoption.The focus of adoption complies with business interest and is in line with the requirement engineering and customer orientation spirit. Through iterative refinement of domain knowledge, we have achieved an accuracy rate of 98,7% in clustering user requirements based on reviews. Our research makes a novel contribution by exploiting machine learning for data-driven innovation in e-commerce. We also lay the groundwork for further research in domains such as individualized content marketing and communication. This method can help e-commerce platforms meet user expectations regarding product innovation more effectively, driving growth and customer satisfaction. In this paper, we,first, carry out a literature review, then we explain our approach and applythis approach to get experiment results by using real data from a company. Finally, we conclude and report future works.

Keywords*—product innovation, marketing, machine learning, requirement engineering, e-commerce*

I. Introduction

The digitalization of the economy is taking place very fast and macroeconomic changes are strongly influenced by the digital aspect (United Nations, 2019). Digital business will grow even more with the Covid-19 pandemic that is spreading throughout the world. In addition, the development of fiber optic networks that continues to expand is an enabler for digital business growth (OECD, 2020). Digital business services are very dependent on their acceptance, the wishes and needs of users [11]. The ability of digital service platforms to provide solutions and provide excellent service to users are the key to the acceptance and use of e-commerce [13], [12]. This is also a determinant of whether users switch to other providers or not depending on satisfaction as a result of the ability of the digital service platform to be able to meet user expectations and desires [7]..A wealth of potentially valuable information about product innovation exists on social media [4], [14]. E-commerce users also share their aspirations and perceptions online [2][3], those papers focus on these users.Because most efforts to take advantage of this data are manual, they cannot handle a large amount of data. One promising way to analyze this data is to employ machine learning [1]. The earlier data of user aspirations and desires are obtained, the earlier the engineering process and the innovation of the e-commerce platform system can be engineered.Speaking about innovation, the innovations is the creation and implementation of new processes, products, services and methods of delivery, which result in significant improvements in outcomes, efficiency, effectiveness or quality [15]. Innovation, which is a key driver of productivity growth, is subject to several well-documented market failures that lead to under-investment in R&D activities [16].The following sub chapter presents a way to identify potential product or service innovations in the e-commerce domain.

II. State of Art

Similar to HYVE project team in Netnography[9], [10], we focus on review and comment possibilities in Internet about products to be able to use the customer feedback in social media and eCommerce for innovation and product politics purposes. In our research, similar to [9], [10], we use the artificial intelligence algorithms to deal with big data, but we go further in direction requirements engineering (RE) for innovation than Netnography's researchers do.

Some researches published results regarding potentially valuable information about product innovation that exists on social media [4], [14]. E-commerce users also share their aspirations and perceptions online [2][3]. Those papers focus on these users. We research on usable data content for innovation.

In the field of requirements engineering one could find many research results like Lim [6]'s researches. They published articles related to an automated and data driven requirement elicitation. Lim [6] analyzed 1848 articles related to an automated and data driven requirement elicitation. He extracted three aspects of the solution step, namely:

- Types of dynamic data sources used for automated requirements elicitation
- Techniques used for automated requirements elicitation
- The outcomes of automated requirements elicitation

Lim found out that there is a clear dominance of human-sourced data, compared to the process-mediated and machine-generated data sources. As a result of that the techniques used for data processing are based on natural language processing, while the use of machine learning for classification and clustering is prevalent. The dominant intention of the proposed methods was to automate the elicitation process fully, rather than to combine it with traditional stakeholder-involved approaches. The final results regarding the completeness and the readiness of the elicited data for use in system development or evolution are currently limited—most of the studies obtain some of the information relevant for requirement's content, some studies target the identification of the core functionality or quality in terms of features, and only a few of the studies achieve a high-level requirement content.

Especially in context of requirement elicitation method Lim identified the following three common steps: (1) filtering out data irrelevant to requirements, (2) classifying text based on the relevance to different stakeholder groups, or (3) classifying text by categories of technical issues, such as bug reports and feature requests. We accomplished this classification using rule-based approaches and machine learning, mostly within the supervised learning paradigm.

Another scholar an aspect for the requirement elicitation, namely an automated prioritization of the requirement. Avesani[8] introduced a framework based on a requirements prioritization process that interleaves human and machine activities, enabling an accurate prioritization of requirements. Fig.1[8] depicts the basic process that the evaluator undertakes. The types of data involved in the process are depicted as rectangles, namely: Requirements represent data in input to the process, that is the finite collection of requirements that have to be ranked; Requirements pair is a pair of candidate requirements whose relative preference is to be specified; Preference is the order relation between two alternative requirements elicited from the stakeholder. The preference is formulated as a boolean choice on a pair; Ranking criteria are a collection of order relations that represents ordering induced by other criteria (e.g. the cost for the realization of the requirements, the estimated utility) defined on the initial set of requirements; Final ranking represents the resulting preference structure on the set of requirements. This final ranking may become the input to a further iteration of the process.

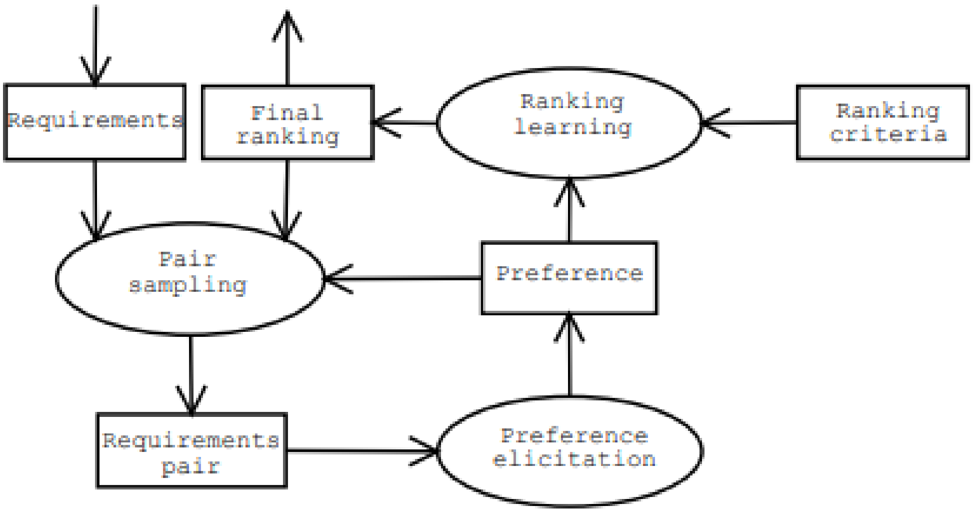


Fig.1:Basic Iteration of Requirements Prioritization Process

We enhanced our published approach of InnoCrowd - an AI Based Optimization of a Crowdsourced Product Development - [5] and we adopted new ideas from formerly published studies to develop a new method for an AI-based semi-automated requirement elicitation for product innovation in data driven marketing and applied it to the e-commerce domain.

III. Our Contribution of AIT-based Requirement Elicitation for Product Innovation

A. *Overview and methodology of our approach*

Fig.2shows the overview of our inter- and trans-disciplinary research approach. Indeed, we combine in this paper the domain of marketing and the domain of requirement engineering.

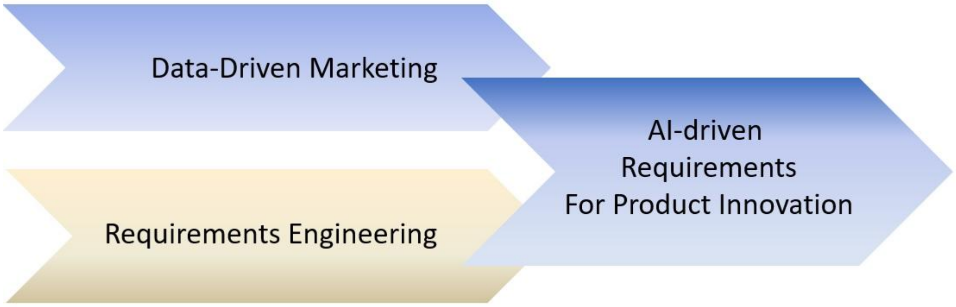


Fig.2 Overview of our interdisciplinary approach

Our approach is based on our results regarding InnoCrowd[5]. For this research, we do not focus only on crowdsourcing but according to data-driven marketing, we consider input of all users (or customers) in ecommerce platforms and use this input as text for machine learning to provide AI-driven requirement elicitation for product innovation, as shown inFig. 3.

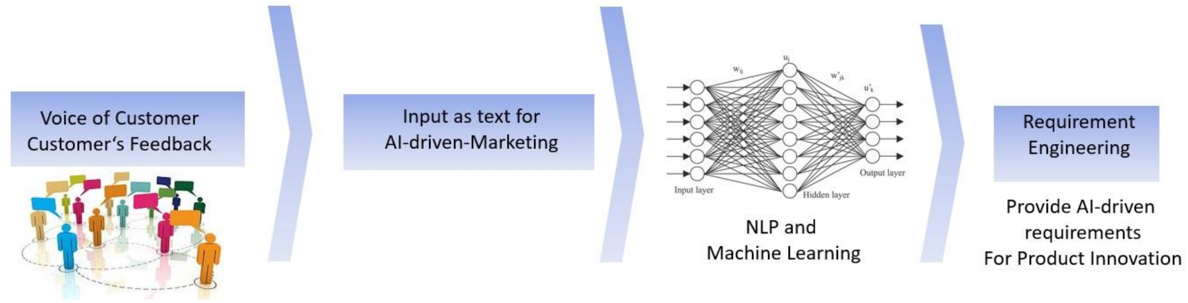


Fig. 3: Approach methodology of AI-based requirement elicitation

B. *Method process of our approach*

We propose to adopt ideas from requirement engineering studies mentioned in state of the art, to consider results regarding machine learning and AI based marketing and to combined them to develop a new method for an AI-based and semi-automated requirement elicitation and applied it to the e-commerce domain. This method will use the base for the data classification related to our two main aspects mentioned in the previous chapter, e-commerce and innovation in e-commerce. Our method consists of the steps depicted in Fig. 4.

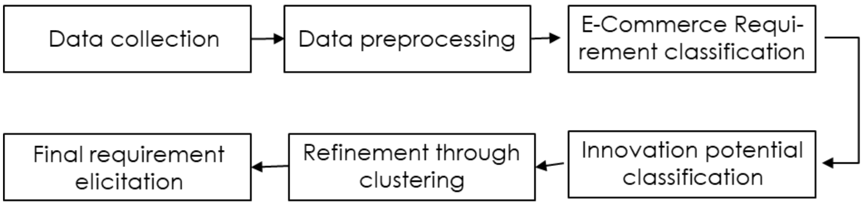


Fig. 4: Method process of new approach

Our method combines the state-of-the-art solutions with new aspects related to innovation, namely on the fourth step “innovation potential classification”. The sixth step “final requirement elicitation” adapts the state of the art to prioritize the requirement and structure it. The next chapter explains each step and the application of our method in detail using a case study.

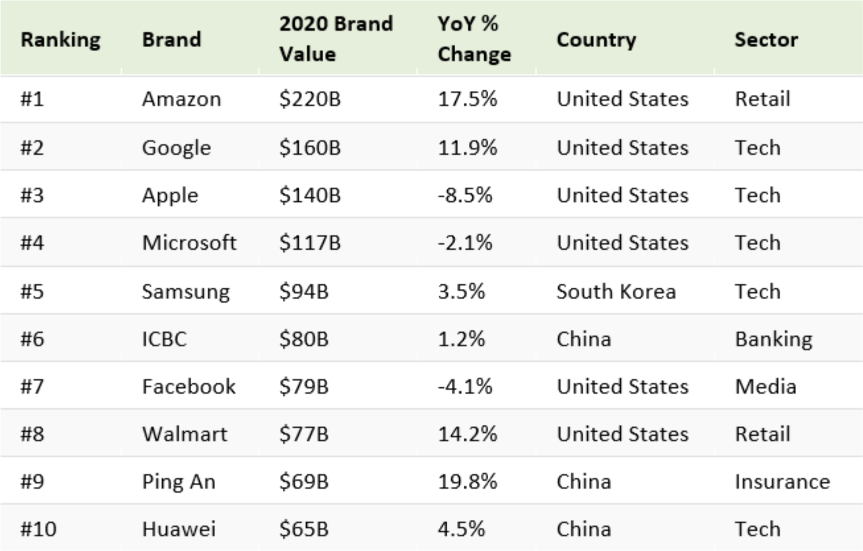
IV. Results of applying our approach on a use study

Schindler & Parent GmbH supports this academic research project in providing the research team use cases. Our use case for the first research paper is a product whose applied results could be of interest for many Schindler & Parent GmbH’s partners. For a generic research in requirements engineering domain that can be applied in many branches, we need requirements that can be turned into specifications. We decided to work with use case product fully automated coffee machine. So, we applied each step of our proposed new method for this product.

First, we collected the data related to e-commerce from scraping tools Mention Lyrics®, TalkWalker®, Octoparse®, and self-written program in Python using Anaconda® as development environment on fully automated coffee machines of Siemens®, Philips® and delonghi®.

At the beginning of the data collection, we needed to define the scope of data related to e-commerce.Table 1shows our starting position for data collection. Table 1shows leading e-commerce companies that together provide a sufficient number of qualitative data to analyze requirements. We get this data from the leading e-commerce companies. This dataset appeared large enough to support our project. After having carried up the first step of our method, we could get a first dataset with 11184 items. For further research we intend to get more items.

TABLE 1: LEADING E-COMMERCE COMPANIES

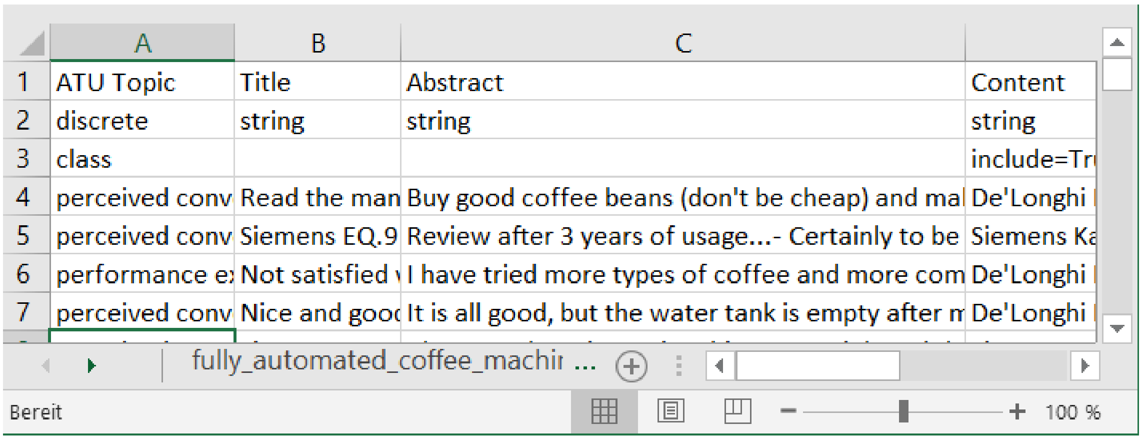


Ranking	Brand	2020 Brand Value	YoY % Change	Country	Sector
#1	Amazon	$220B	17.5%	United States	Retail
#2	Google	$160B	11.9%	United States	Tech
#3	Apple	$140B	-8.5%	United States	Tech
#4	Microsoft	$117B	-2.1%	United States	Tech
#5	Samsung	$94B	3.5%	South Korea	Tech
#6	ICBC	$80B	1.2%	China	Banking
#7	Facebook	$79B	-4.1%	United States	Media
#8	Walmart	$77B	14.2%	United States	Retail
#9	Ping An	$69B	19.8%	China	Insurance
#10	Huawei	$65B	4.5%	China	Tech

b.

All collected items had to be prepared for machine learning training: data cleaning according to machine learning and artificial intelligence process (Machine Learning and Artificial Intelligence 2020) which is commonly called data preprocessing. In this step, we cleaned up the text data (data preprocessing) from any unnecessary text components such as "," "?" "and" "then", which are not relevant for the NLP text analysis. We use the tool Orange data mining® for this preprocessing. Preprocess Text splits the text into smaller units (tokens), filters them, runs normalization, 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 third step used the base of data analysis mentioned in Fig. 4. We used the following aspects: perceived convenience, performance expectation that can be turned into functional and non-functional requirements. During this step, we manually classified the data according to these aspects as seen in Fig. 5.



	A	B	C	
1	ATU Topic	Title	Abstract	Content
2	discrete	string	string	string
3	class			include=Tr
4	perceived conv	Read the man	Buy good coffee beans (don't be cheap) and mal	De'Longhi
5	perceived conv	Siemens EQ.9	Review after 3 years of usage...- Certainly to be	Siemens Ka
6	performance e:	Not satisfied |	I have tried more types of coffee and more com	De'Longhi
7	perceived conv	Nice and good	It is all good, but the water tank is empty after m	De'Longhi

Fig. 5: Exported dataset for machine learning

We separated the dataset into two parts, 80% for the training of the algorithm and 20% for validation / testing of the algorithm. We compared the performance of several machine learning algorithms by using the tool Orange data mining®.

Fig. 6 shows the process chain for the integrated text analysis out of Orange data mining®.

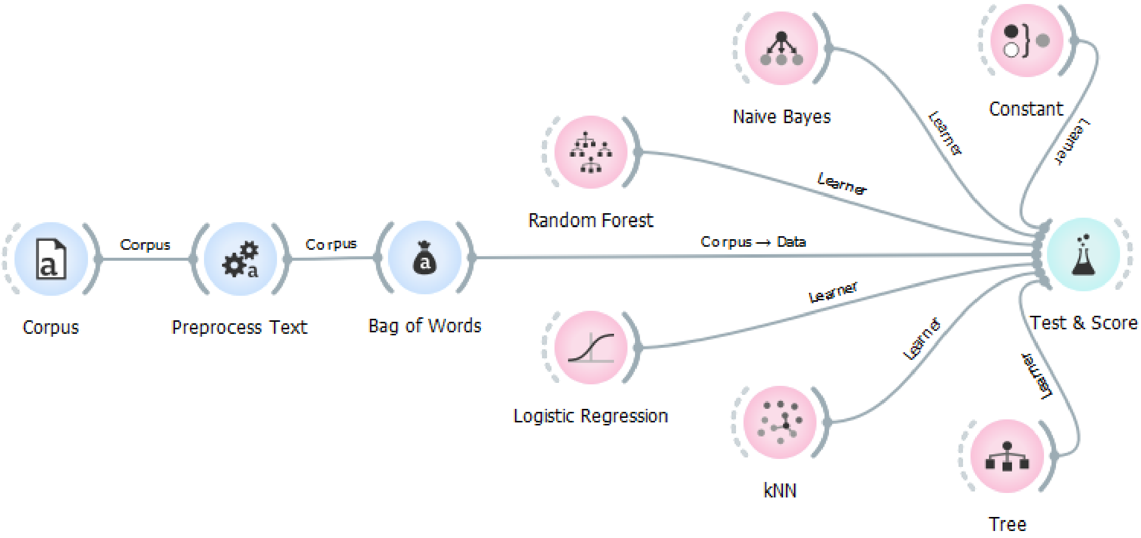
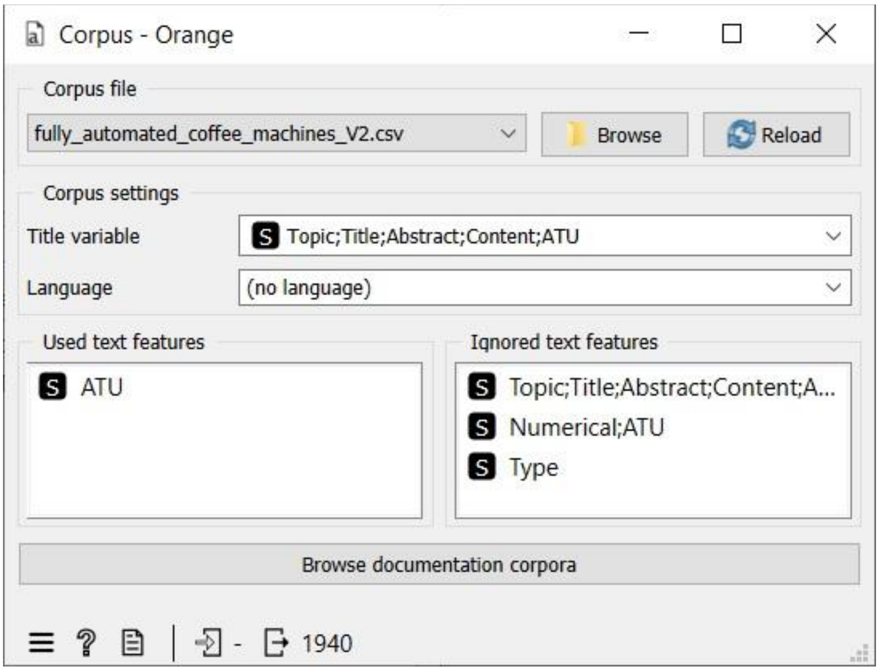


Fig. 6Process chain for the integrated text analysis out of Orange data mining®

In Fig. 6the process chain for the text classification starts with „Corpus", where the text data is uploaded to Orange data mining®tool. In "Corpus", we import the dataset as shown in Fig. 7


Corpus - Orange

Corpus file
fully_automated_coffee_machines_V2.csv
Browse
Reload

Corpus settings
Title variable: S Topic;Title;Abstract;Content;ATU
Language: (no language)

Used text features
S ATU

Ignored text features
S Topic;Title;Abstract;Content;A...
S Numerical;ATU
S Type

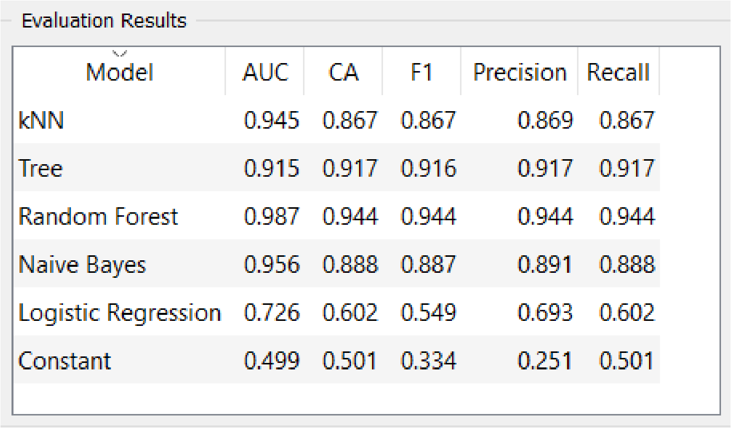
Browse documentation corpora

- 1940

Fig. 7: Dataset uploaded as Corpus in Orange data mining®

The second step is „Preprocess Text" as described in the second step of Fig. 4. The third step „Bag ofWords" is an NLP (Natural Language Processing) mechanism for processing the text so that it can be classified by machine learningalgorithm in the next step. It considers among others the frequency of relevant keywords for each text classification class. During the fourth step we use and compare the performance of six machine learningalgorithm, namely Random Forest, Naive Bayes, Constant, Logistic regression, kNN and Tree. All the text classification result can be found under the step „Test &Score".

We tested andFig.8shows the classification results collected in „Test &Score".



Evaluation Results

Model	AUC	CA	F1	Precision	Recall
kNN	0.945	0.867	0.867	0.869	0.867
Tree	0.915	0.917	0.916	0.917	0.917
Random Forest	0.987	0.944	0.944	0.944	0.944
Naive Bayes	0.956	0.888	0.887	0.891	0.888
Logistic Regression	0.726	0.602	0.549	0.693	0.602
Constant	0.499	0.501	0.334	0.251	0.501

Fig.8: Result of the text classification with several algorithms

We can see that Random Forest delivers the best accuracy on the text classification task: AUC = 0,987, CA = 0,944, Precision = 0,944. Therefore, we use Random Forest for the following text classification task.

During the last (sixth) step, we need to analyze if the machine-based clustering result is meaningful and better than the requirements structure given by the fifth step manually. We have to compare it with the machine-based clustering.

The importance of a requirement can be identified by calculating the frequency of this element being mentioned by the e-commerce users. The importance of a requirement can be identified by calculating the frequency of this element being mentioned by the e-commerce users. It is helpful for e-commerce companies to know which requirements are relevant to their e-commerce product or tool.

V. Conclusion and Future Work

In this paper we present our method for an AI-based semi-automated requirement elicitation based on usable reviews on products in the e-commerce und social media domain by using machine learning for innovation and product politics.

The first results of our research have been published in this paper. This research is not finished. We have to analyze the requirements structure given by the fifth step manually and compare it with the machine-based clustering learnt by comments. (seeFig.8).

In order to get a good data mining result, we need to incorporate the correct domain knowledge. It is based on the state of the art with enhancements of the analysis of e-commerce and innovation aspects to produce valid requirements. We achieved an accuracy of 98,7 %.

To build upon this work, we plan to consider a more complex framework related to the e-commerce and requirements engineering domains to improve machine learning performance. We also need to extend our research work to other marketing domains like individualized content marketing and communication.

Acknowledgment

This research has been supported by Schindler & Parent GmbH, in Meersburg, Germany. We thank Schindler & Parent GmbH for this support.

References

[1] Bohr, A. and Memarzadeh, K., Eds. 2020.*Artificial intelligence in healthcare data*. Academic Press, Amsterdam.

[2] Dwivedi, Y. K., Ismagilova, E., Hughes, D. L., Carlson, J., Filieri, R., Jacobson, J., Jain, V., Karjaluoto, H., Kefi, H., Krishen, A. S., Kumar, V., Rahman, M. M., Raman, R., Rauschnabel, P. A., Rowley, J., Salo, J., Tran, G. A., and Wang, Y. 2021. Setting the future of digital and social media marketing research: Perspectives and research propositions.*International Journal of Information Management* 59, 102168.

[3] Dwivedi, Y. K., Ismagilova, E., Sarker, P., Jeyaraj, A., Jadil, Y., and Hughes, L. 2021. A Meta-Analytic Structural Equation Model for Understanding Social Commerce Adoption.*Information systems frontiers*, 1–17.

[4] Farina, M. 2020. Using Conversation Analysis for Examining Social Media Interactions. In*Machine Learning and Artificial Intelligence*. IOS Press, 172–177.

[5] Kusumah, I., Rohleder, C., and Salinesi, C. 2022. InnoCrowd, An AI Based Optimization of a Crowdsourced Product Development. In . Springer, Cham, 267–278. DOI=10.1007/978-3-030-94335-6_19.

[6] Lim, S., Henriksson, A., and Zdravkovic, J. 2021. Data-Driven Requirements Elicitation: A Systematic Literature Review.*SN COMPUT. SCI.* 2, 1, 1–35.

[7] Mei Ling Goh, Tan Seng Huat, Elaine Ang Hwee Chin, and Mei Qi Yap. 2020. CUSTOMER SATISFACTION AND BRAND SWITCHING INTENTION OF MOBILE SERVICE AMONG UNIVERSITY STUDENTS.*Management & Accounting Review (MAR)* 19, 2.

[8] Paolo Avesani, Cinzia Bazzanella, Anna Perini, and Angelo Susi. 2004. Supporting the Requirements Prioritization Process. A Machine Learning approach. In , 306–311.

[9] Robert V Kozinets. 2010. Netnography: The Marketer's Secret Ingredient How Campbell's and other marketers are using the tools of anthropology online.*MIT's technology review*.

[10] Robert V. Kozinets. 2020. Netnography Today : A Call to Evolve, Embrace, Energize, and Electrify. In*Netnography Unlimited*. Routledge, 3–23. DOI=10.4324/9781003001430-2.

[11] Ruggieri, R., Savastano, M., Scalingi, A., Bala, D., and D’Ascenzo, F. 2018. The impact of Digital Platforms on Business Models: an empirical investigation on innovative start-ups.*Management & Marketing* 13, 4, 1210–1225.

[12] Sharma, G. and Lijuan, W. 2014. Ethical perspectives on e-commerce: an empirical investigation.*Internet Research* 24, 4, 414–435.

[13] Sharma, G. and Lijuan, W. 2015. The effects of online service quality of e-commerce Websites on user satisfaction.*The Electronic Library* 33, 3, 468–485.

[14] Tallón-Ballesteros, A. J. 2020.*Machine Learning and Artificial Intelligence*. *Proceedings of MLIS 2020*. Frontiers in Artificial Intelligence and Applications Ser v.332. IOS Press Incorporated, Erscheinungsort nicht ermittelbar.

[15] Taylor, S. P. 2017. What Is Innovation? A Study of the Definitions, Academic Models and Applicability of Innovation to an Example of Social Housing in England.*JSS* 05, 11, 128–146.

[16] Zhou, L. 2021.*Essays on the Economics of Innovation: Incentives, Diffusion, and Disparity* 10813. EPFL. DOI=10.5075/epfl-thesis-10813.