Improving Product Question Answering in E-Commerce

RICARDO RODRIGUEZ, DETI - Universidade de Aveiro, Portugal

Product Question Answering (PQA) plays a huge role on helping shoppers find the right information to make informed purchasing decisions in E-Commerce. Recent PQA systems consider different approaches to improve generated answers to product-related questions, from considering user-generated content, product reviews and even humor detection.

Additional Key Words and Phrases: PQA, personalization, humor, answer, question, review, product

1 INTRODUCTION

Product Question Answering (PQA) is a task on the natural language processing domain that involves answering questions about products, such as their features, usage and specifications. *"Is this camera model better than Y?"*, *"are the headphones compatible to Bluetooth 4.0?"* and *"for how long does this HEPA filter works without replacement?"* are examples of some questions users can ask to get information on a specific product.

Therefore, PQA plays a crucial role in e-commerce by helping customers find the information they need to make informed and confident purchase decisions. Providing accurate and timely responses has a significant influence on the customer decision to buy a particular product, on reducing the workload of customer service departments and on improving customer satisfaction with the platform.

Given that there is significant revenue generated by e-commerce platforms, with retail e-commerce sales constituting approximately 4.9 trillion U.S. dollars [2], it is important for these companies to improve PQA systems in order to enhance the customer experience and increase sales.

In this monograph, we will explore the state of the art of PQA systems which aim to improve the customer experience in e-commerce platforms. We will examine some interesting approaches used to generate answers to user questions about products and evaluate their effectiveness and limitations. In the end, we will draw some conclusion on the studied methods and discuss the potential future directions of this field.

2 STATE OF THE ART

Many different methods have been implemented to improve product-related question answering, from using knowledge of similar products [6], using semi-structured answer sources [9], retrieving information from multiple sources [8] and considering relevant reviews as candidate answers by measuring its similarity to the given question [11].

Although these approaches to PQA are effective and improve user experience towards confident product purchase, they generate the same answer for questions made by different individuals, neglecting an important characteristic of human nature: personality.

According to Deng, Y. et al. (2021), the majority of user-generated questions are subjective (56.8%) rather than objective. However, 80.9% of answers reflect personalized information (*e.g.* personal experience, preferred product aspects), since almost all helpful answers to subjective questions involved personalized information and half of the helpful answers to objective questions still contain personalized information. We can observe two different types of user preference for those biased answers: users worried with personal experience and users worried with product-level aspects. These observations were taken by conducting a statistical user study by analyzing the Amazon QA datasets, which represent real-life e-commerce data, represented in *Fig. 1*. [13].

Thus, the importance of personalization in a PQA system it's undeniable and the generated answers should also reflect the interests and preferences of the user and cut out his/her concerns on a specific product.
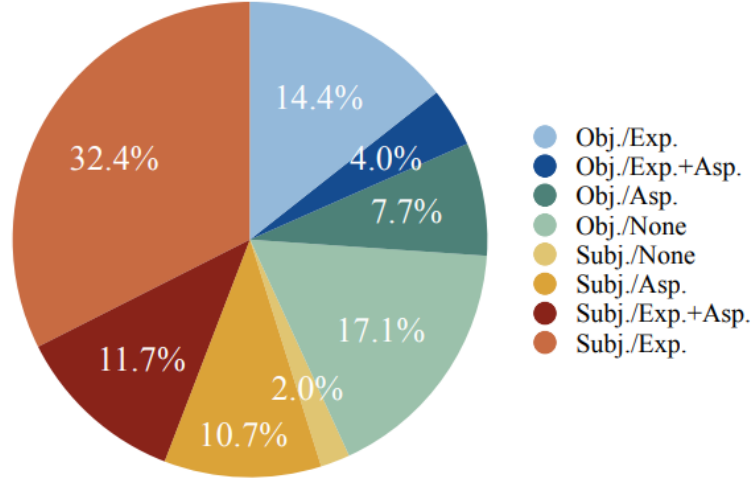

Fig. 1. The Statistics of User-generated Questions and Answers on Amazon QA (best viewed in color). Subj. and Obj. denote subjective and objective questions respectively. Exp. and Asp. denote the answers with personal experience or preferred product aspects, respectively, while None denotes the answers that cannot reflect any user preference. [13]

Additionally, different costumers have different preferences and concerns over product characteristics, such as price, quality, weight, compatibility and a wide range of aspect-level information which should take into consideration when generating an answer to a certain individual. To achieve this, comprehensive user preference modeling is necessary, rather than providing a uniform response to the same question [13].

One way to integrate the user personalization component into PQA systems is by using user-generated content (UGC), such as reviews, answers and questions, which is widely available on e-commerce platforms. UGC provides valuable insights into how customers use and perceive products, including their product-level preferences. By taking into account user personalization, PQA systems can provide more relevant and accurate answers to customer questions.

To tackle this problem, Deng, Y. et al. (2021) proposed Personalized Answer GEneration method (PAGE), a PQA system that analyzes relevant user-generated content to the given question to model user preference at knowledge-, aspect- and vocabulary-level. It's comprised of four modules:

- *Base Encoder-decoder Architecture* is the base model of PAGE. It uses the question X_q and the relevant supporting facts X_f as input and outputs the vocabulary probability distribution $P_v(a)$, not considering any type of user personalization.

- *Persona History Incorporation* (PHI) which retrieves user-generated content which outputs the persona history memory vector M_h for knowledge-level preference modelling. Thus, it's possible to understand what customers expect to receive according to their own experience, which can be discovered from UGC.

- *Persona Preference Modeling* (PPM) module employs neural topic model to capture latent aspect- and vocabulary-level user preference. Knowledge-level modeling of user preference is not sufficient to generate personalized answers for a specific customer. Hence, it's important to capture user preferences in relation to product characteristics (aspect-level) and reflect user vocabulary (vocabulary-level) to provide better customer experience and increase e-commerce sales.

- *Persona Information Summarizer* (PIS) module summarizes the multi-perspective user preference information for generating personalized answers, which aggregates the knowledge-level information produced by the PHI component and the aspect- and vocabulary-level received from the PPM component to create a persona-aware pointer generator network.

The architecture of the PAGE method and its main components, mentioned before, are depicted in *Fig. 2*. In order to fully understand the proposed PQA system, we suggest reading the paper written by Deng, Y. et al. (2021) [13].
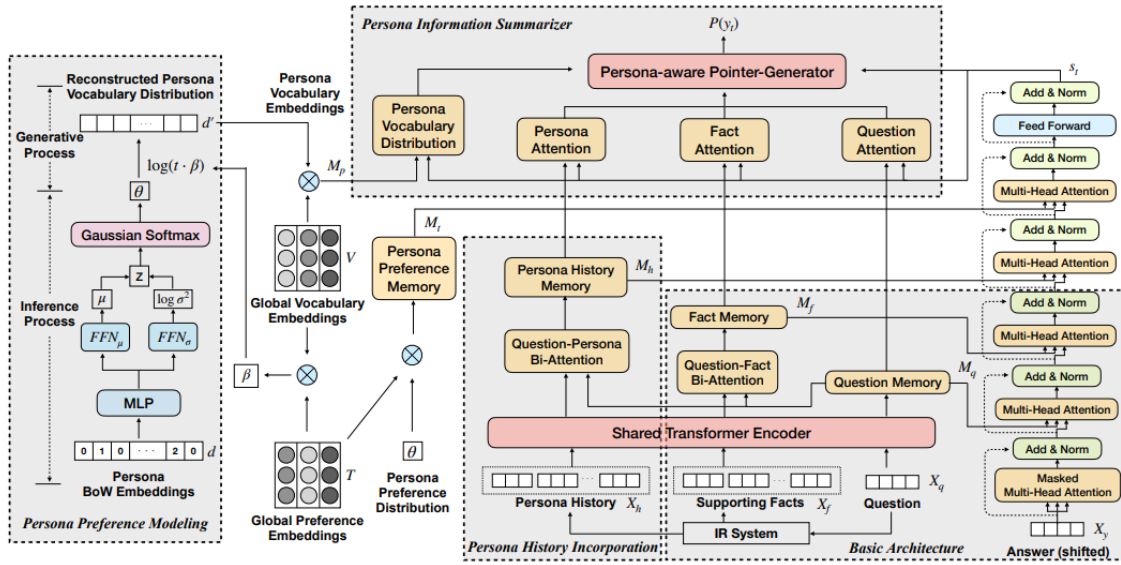

Fig. 2. Overview of the proposed method PAGE, including four components: Basic Encoder-decoder Architecture, Persona History Incorporation, Persona Preference Modeling, and Persona Information Summarizer [13]

Even though product question answering is greatly improved with the multi-perspective preference modeling solution proposed by Deng, Y. et al. (2021), which explores UGC to model user preference for generating personalized answers in PQA, there's also an important factor to consider: humor.

E-Commerce platforms with community question-answering (CQA) allow users to post questions and get answers from the community in relation to a certain product. Managing the content quality of these PQA systems is a difficult task, as some questions may be more relevant to many users, while others may be removed for being offensive, confusing or simply false. Humorous questions, in particular, can be a double-edged sword since they can either be viewed as engaging to some users but confusing or even offensive to others. Some examples of differences between humorous and non-humorous questions can be observed in *Fig. 3*, alongside the specific product that received the respective question.

Product	Image	Humorous Questions	Non-Humorous Questions
Nintendo Switch Gray Joy-Con		• Can i use this to hack into the matrix and save humanity? • Can I trade one of my kidneys? • What if the princess wants to be with Bowser and Mario keeps kidnapping her?	• What do the ports on the side of the console do? • How much money will the system cost? • How do I know if this is the neon or gray version?
Echo Show - 1st Generation Black		• Will this help me find the meaning of life? • Can Alexa show me my future? • Does it cook breakfast?	• Can you see YouTube videos? • Can you see your Echo Show camera on the cloud app? • Can it connect to music speakers?
Sovaro Luxury Cooler		• Will this thing make me fly? It seems due to the price that it has to do something special • Which organ should I sell to finance this ice box? • Just how insecure do you have to be to buy one?	• Will this fit in the trunk of my Lambo? • Where do you plug it in? • What is the country of origin?
Hutzler 571 Banana Slicer		• I set it down in my kitchen, my bananas have stopped talking to me. What now? • What if the banana bends the other direction? • Is there a model for left-handed people?	• Can this be used on cucumbers? • Does it only come in yellow? • Seriously, though - does this thing really work?

Fig. 3. Example of products associated with humorous and non-humorous questions [1]

Previous studies have shown that humor detection accuracy can be have major improvements when text is associated with context [4]. On the other hand, lack of context complicates the task of identifying humor in text sentences. However, popular PQA systems can benefit from product details, as they can act as context to a given question. Quoting an example from Amazon, Y.Z. et al. (2020), the question "can it also be used for making coffee?" has a non-humorous intent when asked in the context of a tea pot and a humorous intent in the context of a Swiss army knife. Furthermore, there are unique text properties that can be used to assist humor detection on customer questions: incongruity, which describes the contradiction between the question and the product (e.g. *"Can I dress this TV?"*), and subjectivity, when a customer asks a question that somehow reflects subjective opinion (e.g. *"Will this TV make me happy?"*) and may vary from individual to individual, as previously mentioned when analyzing Deng, Y. et al. (2021) work. Moreover, one of the main obstacles to humor detection is domain bias, which happens when the system tries to detect humor in text according to its domain. For example, adult toys, masks, bizarre products and too expensive products often attract more humorous questions than vulgar products (e.g. electronics, clothes, books). Consequently, the classifier may be trained to identify these products instead of identifying humorous questions [1].

In order to solve these problems, Amazon, Y.Z. et al. (2020) created a PQA system with a deep-learning humor detection framework, capable of identifying the differences between non-humorous and humorous questions to improve user engagement with the e-commerce platform. Overall, the system retrieves information on the question and gathers product details to build context. Additionally, the module analyzes the incongruity between the question and the product-built context and identifies subjective questioning. According to Amazon, Y.Z. et al. (2020), in order to eliminate the product bias, the authors experiment with two balanced datasets, both sharing the same set of humorous questions, which are different in the way negative examples are selected. For the biased dataset, negative examples are selected at

random from the entire population of questions. For the unbiased dataset, a negative example is selected at random, for each positive question, from the set of non-humorous questions that match the question's matching product. As expected, a classifier trained over the unbiased dataset, achieves lower accuracy than a classifier trained over the biased dataset. However, the unbiased classifier excels in the task of differentiating between humorous and non-humorous questions that match the same product [1].

The workflow of the humor detection framework proposed by Amazon, Y.Z. et al. (2020) [1] is depicted in *Fig. 4*.
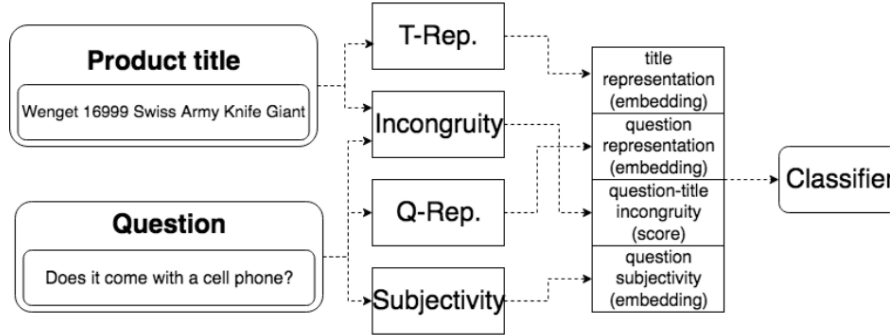

Fig. 4. Example of the workflow of the humor detection framework, given a question and an associated product [1]

As we can observe from *Fig. 4*, the framework starts by taking a question and a specific product as input and includes two pre-trained modules that capture incongruity and subjectivity, which have a major role on capturing humor in text, and the $T - Rep.$ and $Q - Rep.$ representational models that capture text features from the product title and the question, respectively. In the end, the results of the modules are concatenated to form the final vector, comprised by the title representation, question representation, question-title incongruity and question subjectivity single vectors, which serves as an input to the classifier that produces the humorous result. It's important to note that to represent the product title and the question, a Long Short-Term Memory (LSTM) is used to capture the inner features of text, being the last hidden layer of LSTM the vector representation. Questions highly coupled to a certain product are considered to have low incongruity and, thus, are less likely to be humorous, while a dissimilar relationship between a question and a product are considered to have high incongruity and, thus, more likely to be taken as a humorous question. To find if the two text snippets are associated to each other, a Siamese network is implemented [7]. Finally, the subjectivity module is very important since some customers make humorous questions which express an individual trait (e.g. sentiment, opinion, etc.). As mentioned by Amazon, Y.Z. et al. (2020), the question "Will this thing make me fly? It seems due to the price that it has to do something special" towards a luxury cooler clearly expresses discontentment about the price of the product. Concerning sentiment polarity analysis, whose goal is the correct classification of positive and negative sentiments in texts, the framework authors trained an LSTM model based on the Blitzer's dataset, which contain Amazon product reviews [3]. The sentiment of each review is labeled according to its associated human star rating. However, the authors found the sentiment polarity signal was unhelpful in detecting humorous questions, since an initial examination showed an accuracy of 52% in humor detection.

Although the previous two methods focus on the different aspects of each individual, their previously generated content and the detection of emotions when asking questions on a specific product, E-commerce platforms still have the problem of users having to browse through extensive customer reviews and community question-answering systems

(CQA) to find information about a product when shopping online. These systems tend to have a low recall rate and users typically have to wait an extended period for the questions to be responded to. Additionally, extracting relevant information from reviews can be difficult due to syntax errors and lack of punctuation.

To address this issue, Chen et al. (2019) proposed Review-driven framework for Answer Generation in E-commerce (RAGE) [5], a solution that uses a multi-layer convolutional network architecture [12] to speed up answer generation with parallel computation, as opposed to other PQA systems which use recurrent neural networks (RNN) [10]. The workflow of the RAGE system is illustrated in *Fig. X*.
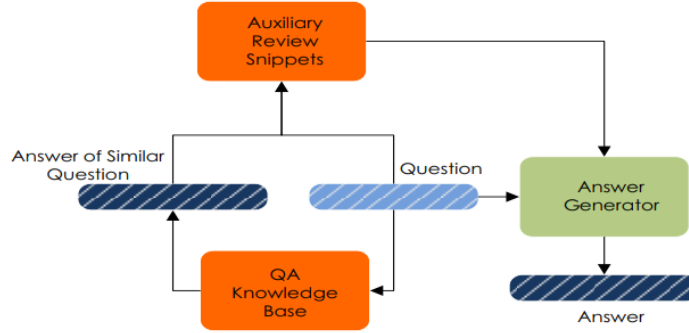

Fig. 5. Workflow of the RAGE proposed system [5]

The overall architecture of RAGE, depicted in *Fig. 4* and fully explained in the paper written by Chen et al. (2019) [5], can be divided into two different modules:
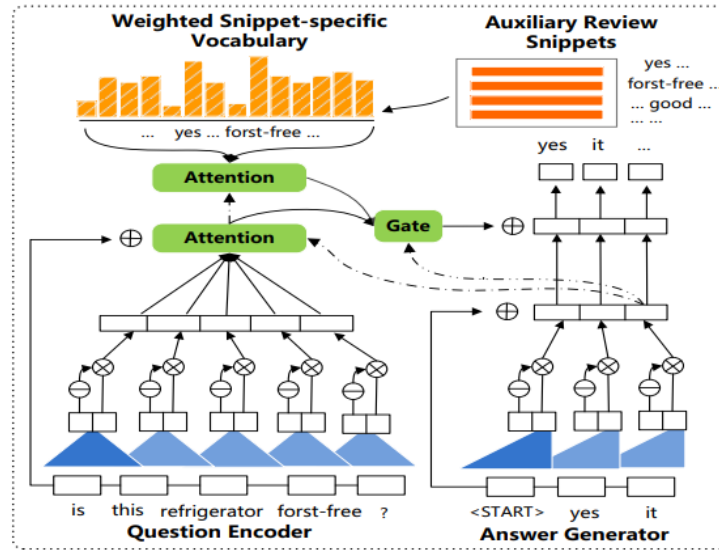

Fig. 6. Overall architecture of the RAGE proposed system [5]

- *Question Encoder* module is responsible for constructing the dense vector representation of each word of the question. The component uses a sliding window of size k to extract the hidden state for each word with the help of a multi-layer gated convolutional network. This deep learning algorithm stacks layers on top of each other to create a hierarchical structure over the whole question, noting that the higher layers consider more distant words through the hidden states encoded by the tower layers, which connect both consecutive layers. At the last layer (output layer), a linear transformation is applied to the hidden states, which serve as an output to the question encoder component.

- *Review-Driven Answer Generator* module acts as the Question Encoder one, using a convolutional network for answer generation. Chen et al. (2019) introduced a special symbol $y0$ indicating the start of the answer and, after generating j-th words by step j, we have an answer sequence $T = (y0, y1, ..., yj)$. The hidden state for word y_j at each layer is hierarchically calculated through a formula [5] and, at the the output layer, the generation probability of word $y_j + 1$ given the j previous words is calculated through a linear transformation and a softmax function that takes the hidden state calculated at the last layer for word j and the parameter matrix W and bias vector b.

For each given question, the RAGE system extracts the auxiliary review snippets that contain relevant information from the customer reviews of a specific product. However, each one of these review examples could contain noisy information that could potentially harm the system. We can assume review tokens are of major importance and represent the core information about the product when they dominate the review snippets. Thus, the proposed system calculates the weight of each word present in the snippet reviews according to frequency and semantic relatedness, making irrelevant words obsolete and reinforcing core tokens. After this, the token weights are normalized by using max-normalization and the attention and gate mechanisms are utilized to inject relevant review snippets information into the answer generator module.

3 SYSTEMS RESULTS AND DISCUSSION

Even though an evaluation method that compares all the three previously mentioned approaches to the PQA domain in E-Commerce doesn't exist, only existing the comparison between the PAGE [13] and RAGE [5] methods, the results of each system still can be analyzed and discussed.

As we can observe from *Fig. 7*, the PAGE method clearly outperforms the RAGE method on answer generation and other baseline methods by a considerable margin. This figure also demonstrates that the RAGE method is inefficient on several evaluation metrics [13].

Model	R-1	R-2	R-L	ES-Ext.	ES-Gre.	ES-Ave.
BM25	13.2	1.8	11.7	35.4	66.6	84.7
S2SAR [23]	13.5	2.3	12.5	35.8	65.6	78.9
RAGE [10]	13.2	2.1	12.0	35.6	65.5	80.2
TFMR [47]	14.8	2.5	13.2	36.1	65.1	_85.2_
Per-S2SAR [66]	14.0	2.5	13.0	35.6	65.5	82.6
Per-CVAE [54]	13.3	2.0	11.9	35.1	63.4	85.0
PAGen [59]	13.5	2.1	12.0	35.5	64.4	84.6
Per-TFMR	_15.1_	_2.7_	_13.5_	36.3	66.9	85.0
KOBE [9]	15.0	2.5	13.3	_36.8_	_67.4_	84.9
PAGE	**16.9**†	**3.5**†	**15.1**†	38.3†	67.7†	85.0
- w/o PHI	16.2	3.3	14.5	37.5	67.1	85.1
- w/o PPM	16.3	3.5	14.6	**38.6**	**67.9**	**85.3**
- w/o PIS	15.7	2.8	13.9	37.1	66.6	85.2

Fig. 7. Method comparisons of Answer Generation on an *Electronics* dataset. † indicates that the model is better than the best performance of baseline methods (underline scores) with statistical significance (measured by significance test at $p < 0.05$ [13]

Since the RAGE method only relies on community question answering (CQA) and auxiliary review snippets to generate natural answers for product-related questions, the quality of the answers depends also depend on the quality and trustworthiness of user reviews and QA to be accurate. Additionally, the efficiency of the RAGE method is deeply harmed when there is a lack of CQA and user reviews since fewer resources mean it's more difficult to find relevant information. However, the RAGE method tackled the problem of merging noisy and unstructured data from CQA and user reviews to identify relevant information and, hence, generate more accurate answers to e-commerce customers.

In regards to the PAGE method, which generates personalized answers via multi-perspective preference modeling, the results are fairly impressive. The PAGE method captures user preference, previous knowledge and vocabulary-level from UGC to model different preference perspectives according to a given customer. According to Deng, Y. et al. (2021), the PAGE method consistently and substantially outperforms other PQA systems on different datasets (e.g. *Electronics, HomeKitchen, SportsOutdoors*, which proves that PQA can benefit from answer personalization, it generates personalized answers with a higher diversity of user-centric information as well as user-preferred language styles, the multi-perspective preference modeling contributes to the personalization of PQA from different perspectives and aids in generating answers with remarkable content quality and, finally, the answers generated by PAGE preserve a higher degree of informativeness, diversity as well as explicitly reflect some user-centric persona, which contributes to a higher overall quality. Despite the fact that PAGE is a very good PQA method that includes personalization, there is room for improvement, such as the retrieval of other types of user data other than historical UGC.

Regarding the PQA system integrated with the human detection framework presented by Amazon, Y.Z. et al. (2020), we can see the performance comparison between the proposed method and other baseline methods in *Fig. 8.*

Configuration	Method	unbiased	biased
Baseline Methods	Logistic Regression	82.11	87.21
	Naive Bayes	81.64	87.62
	$LSTM_Q$	83.52	88.59
	CNN_Q	83.26	88.69
Partial Configuration	$LSTM_{Q+T}$	83.42	89.74
	CNN_{Q+T}	83.63	89.58
	$LSTM\_INC_{Q+T}$	83.58	89.8
	CNN_INC_{Q+T}	83.26	90.34
	$LSTM\_SUB_{Q+T}$	83.81	90.08
	CNN_SUB_{Q+T}	83.71	90.23
Complete Framework	$LSTM\_INC\_SUB_{Q+T}$	**84.41***	90.26
	$CNN\_INC\_SUB_{Q+T}$	84.13	**90.76***

Fig. 8. Accuracy of humor detectors over the two datasets. Statistically significance improvement with respect to baseline methods is marked by '*' , using McNemar test with p-value < 0.05 [1].

Analyzing *Fig. 8*, we can verify that the complete framework outperforms baseline methods and partial configurations by a small margin. However, it's important to note that every partial configuration of the proposed method still has better accuracy results than the baseline methods.

In addition, an experiment on product bias in PQA data, whose results are shown in *Fig. 9*, reveal that we reach an accuracy of 90.76% and 84.41% over the datasets, with and without the bias respectively. This constitutes an improvement of 18.3% and 5.4% in relative error reduction over baseline methods described in previous studies.

Dataset	*unbiasedQClassifier*	*biasedQClassifier*
biased	82.85	90.76
unbiased	84.41	76.21

Fig. 9. Classifiers accuracy over the two datasets (biased and unbiased) [1].

4 CONCLUSION

In conclusion, this study has highlighted the importance of improving Product Question Answering (PQA) systems in e-commerce, as it can increase sales and improve customer satisfaction.

By comparing the three different approaches to PQA, all the methods have different approaches to the optimization of generated answers. However, we found out that the PAGE method outperforms the RAGE method in all evaluation metrics by a large margin, which can be explained since PAGE is more complex and takes different parameters that focus on the individuality of human nature rather than focusing only on user reviews and community questions and answers.

As indicated by the authors of the proposed PQA systems, there is still room for improvement further research is needed to develop more sophisticated methods.

In the future, e-commerce can be substantially improved by PQA systems that include different aspects of the spectrum and use the powers of natural language processing to ultimately provide the best possible customer experience. Furthermore, the growing rate of user data can provide valuable insights into customer preferences, vocabulary and behaviors can, consequently, be collected to generate even more personalized and relevant answers to the individual.

REFERENCES

[1] Amazon, Y.Z. et al. (2020) Humor detection in product question answering systems: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Conferences. Available at: https://dl.acm.org/doi/10.1145/3397271.3401077.

[2] Asendia (no date) E-commerce sales worldwide set to increase by 50% by 2025, reaching 7.5 trillion dollars., Asendia. Available at: https://www.asendia.com/asendia-insights/e-commerce-sales-worldwide-set-to-increase-by-50-by-2025.

[3] Blitzer, J., Dredze, M. and Pereira, F. (no date) Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification, ACL Anthology. Available at: https://aclanthology.org/P07-1056/. Chen, P.-Y. and Soo, V.-W. (no date) Humor recognition using Deep Learning, ACL Anthology. Available at: https://aclanthology.org/N18-2018/.

[4] Chen, P.-Y. and Soo, V.-W. (no date) Humor recognition using Deep Learning, ACL Anthology. Available at: https://aclanthology.org/N18-2018/.

[5] Chen, S. et al. (2019) Review-driven answer generation for product-related questions in e-commerce, arXiv.org. Available at: https://arxiv.org/abs/1905.01994.

[6] Mejer, A. (2021) Predicting answers to product questions using similar products, Amazon Science. Amazon Science. Available at: https://www.amazon.science/blog/predicting-answers-to-product-questions-using-similar-products.

[7] Mueller, J. and Thyagarajan, A. (no date) Siamese recurrent architectures for learning sentence similarity, Proceedings of the AAAI Conference on Artificial Intelligence. Available at: https://ojs.aaai.org/index.php/AAAI/article/view/10350.

[8] Rajasekar, A.A. and Garera, N. (2021) Answer generation for questions with multiple information sources in e-commerce, arXiv.org. Available at: https://arxiv.org/abs/2111.14003. Review-driven answer generation for product-related questions in e-commerce (no date). Available at: https://arxiv.org/pdf/1905.01994v1.pdf.

[9] Shen et al. (2022) semiPQA: A Study on Product Question Answering over Semi-structured Data. Available at: https://aclanthology.org/2022.ecnlp-1.14.

[10] Sherstinsky, A. (2021) Fundamentals of Recurrent Neural Network (RNN) and long short-term memory (LSTM) network, arXiv.org. Available at: https://arxiv.org/abs/1808.03314.

[11] University, L.C.N. et al. (2019) Answer identification from product reviews for user questions by multi-task attentive networks: Proceedings of the Thirty-third AAAI conference on artificial intelligence and thirty-first innovative applications of Artificial Intelligence Conference and Ninth AAAI symposium on educational advances in artificial intelligence, Guide Proceedings. Available at: https://dl.acm.org/doi/abs/10.1609/aaai.v33i01.330145.

[12] Yamashita, R. et al. (2018) Convolutional Neural Networks: An overview and application in radiology - insights into imaging, SpringerOpen. Springer Berlin Heidelberg. Available at: https://insightsimaging.springeropen.com/articles/10.1007/s13244-018-0639-9.

[13] Yang Deng The Chinese University of Hong Kong et al. (2022) Toward personalized answer generation in e-commerce via multi-perspective preference modeling, ACM Transactions on Information Systems. Available at: https://dl.acm.org/doi/10.1145/3507782.