6-18-2022

Deep Learning Enabled Consumer Research for Product Development

Sven Stahlmann
University of Cologne, stahlmann@wim.uni-koeln.de

Oliver Ettrich
Kühne Logistics University, oliver.ettrich@the-klu.org

Detlef Schoder
University of Cologne, schoder@wim.uni-koeln.de

DEEP LEARNING ENABLED CONSUMER RESEARCH FOR PRODUCT DEVELOPMENT

Research in Progress

Sven Stahlmann, University of Cologne, Cologne, Germany, stahlmann@wim.uni-koeln.de

Oliver Ettrich, Kühne Logistics University, Hamburg, Germany, oliver.ettrich@the-klu.org

Detlef Schoder, University of Cologne, Cologne, Germany, schoder@wim.uni-koeln.de

Abstract

"Needmining" is the analysis of user-generated content as a new source of customer needs, which are an important factor in new product development processes. Current approaches use supervised machine learning to condense large datasets by performing binary classification to separate informative content (needs) from uninformative content (no needs). This study introduces a transformer model and compares it to relevant approaches from the literature. We train the models on data composed from a single product category. Subsequently, we test the models' ability to detect needs in a validation sample containing product categories not present in the training set, i.e. "out-of-category" prediction. Our cross-validated results suggest that, based on the F1-score, the transformer model outperforms previous approaches at both in-category and out-of-category predictions. This suggests that transformers can make needmining more relevant in practice by improving the efficiency of the needmining process by reducing the resources needed for data preparation.

Keywords: Deep Learning, Natural Language Processing, Customer Needs, Product Development, Innovation, Machine Learning.

1 Introduction

Competition and increasingly demanding consumers force companies to put substantial resources into finding the next new product. Despite advancements in product development techniques for new market entries, some industries report failure rates between 35% and 45% within the first years of market introduction (Castellion and Markham, 2013). One key to the success of new products is customer needs (Kuehl et al., 2016). Customer needs are the customer's expression of the benefits expected to be fulfilled by a product or service (Timoshenko and Hauser, 2019). To fulfill these, products and services must consist of attributes that address the desired benefits (Timoshenko and Hauser, 2019). Besides identifying these needs, a successful product must also consist of the right mix of the right attributes (Griffin and Hauser, 1993; Krishnan and Ulrich, 2001). In the innovation literature, lead user theory (von Hippel, 1986) discusses involving customers to aid innovation and product development. In the context of crowdsourcing ideas, consumer information can benefit the idea selection process (Hoornaert et al., 2017) and contribute to innovativeness (Roberts and Candi, 2014). These studies suggest that the inclusion of users, or in other words the voice of the customer (see Griffin and Hauser, 1993), in the ideation phase contributes to the overall success of the final product.

The current state of global connectivity through the internet provides marketers with a new perspective on consumer insights: digital user-generated content (UGC). User-generated content comes in many forms. Microblogs like Twitter, comments on Youtube videos, discussions on social media platforms, and almost all major retailers offer customer reviews on their e-commerce platforms. Review sections are particularly relevant for customers who plan to purchase products both on- and offline (Kannan, 2017). Many reviews contain detailed information about product specifications and use-cases directly reflecting customers' needs and wants for products. Therefore, we focus on the value of customer reviews as a subsection of user-generated content.

For companies, such extensive "databases" of freely available voice of customer could be an opportunity to optimize their product development processes to reduce failure rates. Yet, looking at the development of product development processes throughout the years little seems to have changed. Despite the increased access to consumers through UGC, most companies still rely heavily on internal resources. Large multinational corporations are more likely to consult their research and development departments than to consult external sources. They collaborate with expert panels or recruit consumer testers for quality assurance and product reviews, an alternative source of customer insights (Bashir et al., 2017). However, it is also evident that large companies, but especially smaller ones, turn to UGC as a source of "informal consultancy" (Bashir et al., 2017).

Using UGC as a source of consumer needs is not trivial. The amount of available data is difficult to process manually, resulting in only reviewing small samples and, therefore, potentially missing crucial needs not represented. Moreover, many reviews are repetitive or irrelevant, which further reduces the efficiency of manual analyses (Timoshenko and Hauser, 2019). This calls for automated solutions to extract needs from UGC for product development.

Utilizing digital consumer opinions is not a new idea. In the recent past, researchers used consumer comments to analyze the service quality of airlines (Misopoulos et al., 2014), to automatically identify consumer needs with machine learning (Kuehl et al., 2016), and to extract consumer needs in a human-machine hybrid approach (Timoshenko and Hauser, 2019). The main goal in the latter two studies is to separate informative user-generated content from uninformative content using supervised machine learning methods. These studies share that the model can only analyze content from a single product category. This means that each application of needmining to a new product category requires manual labeling and training of a new classification model, both of which take up significant resources.

We plan to close this research gap by developing an alternative approach that can generalize across categories. This would further reduce the cost and time of needs identification and therefore increase the efficiency of the method. In recent years tremendous advancements have been made in the area of natural language processing (NLP) (Talmor et al., 2019). In particular, pre-trained Transformer models, such as RoBERTa (Liu et al., 2019) were able to break several benchmarks in different NLP tasks. These models are promising candidates in the present binary classification problem, i.e. separating user-generated content into being informative (needs included) vs. uninformative (no needs included). Thus we ask: "How do pre-trained transformer-based models compare to current models for need identification in user-generated content?" (RQ1). Additionally, we explore the ability of models to predict across multiple product categories, effectively reducing the number of labeled training data by asking "How do models perform on categories that are different from the category they were trained on?" (RQ2).

To answer these research questions, we employ the RoBERTa model for need identification in customer reviews as a subcategory of UGC. We test this model against all relevant models currently found in the needmining literature, i.e. support-vector-machines (SVM) (Kuehl et al., 2016) and convolutional neural networks (CNN) (Timoshenko and Hauser, 2019). We evaluate the models using two datasets that differ in their composition of included product categories. One contains reviews only from the category present in the training set of the model ("in-category"). The other is evenly composed of 24 product categories not present in the training set ("out-of-category").

Answering our research questions results in two main contributions. First, we provide an overview of model performance for the identification of needs in UGC and test the performance of transformers in the specific context of needmining. Second, we determine how well the models predict across multiple product categories. Overall, this leads to a potential increase in product development efficiency by eliminating the recurring cost of data preparation and machine runtime for performing needmining on products across several categories.

2 Theoretical Background

2.1 Consumer needs and product development

This study draws from three major streams of literature: New product development, marketing, and natural language processing. This allows us to create an information system-enabled approach to customer needs analysis from digital content to improve product development efficiency and generalizability.

The integration of users in innovation and product development processes is by no means a new concept. Whole branches in innovation and product development literature address this topic with different approaches. (von Hippel, 1986) lead user theory proposes finding users who experience a need before the market and integrating them into new product development. Moreover, open innovation (Chesbrough, 2003) suggests opening the traditionally firm-centric research and development to external sources. Empirically, both approaches improve the firm innovativeness (Lilien et al., 2002; Carlsson et al., 2011). All approaches have in common that they help firms identify and understand customer needs, which increases innovativeness and thereby new product success.

As mentioned, the most recent approach to identifying customer needs involves massive, freely available sources of data from the internet. All sorts of demographics contribute to self-motivated content on different platforms such as blogs, review pages, social media, and more. An early study by (Lee and Bradlow, 2011) investigates methods to automate marketing research with online reviews. Instead of relying on NLP, they use a logical assignment approach to among others find relevant product attributes in reviews. They conclude that customer reviews help identify product attributes that traditional methods do not find. Moreover, reviews ratings contribute to selecting the right attributes as sentiment appears to indicate importance.

More recent studies dealing with user-generated content tend to apply machine learning methods, especially within the field of NLP. Misopoulos et al., (2014) use comments from Twitter micro-blogs to analyze the service quality of four different airlines. Their approach makes use of netnography (a methodology specifically developed for a social media context) and sentiment analysis to identify positive and negative comments about their services. From this, they were able to extract specific customer needs for the airline industry.

Similarly, Kuehl et al. (2016) and Christensen et al., (2017) use Twitter micro-blog and Lego user Forum data respectively to identify customer needs. The main difference to Misopoulos et al. (2014) is the use of a machine learning approach over a sentiment analysis to identify the existence of a need within user-generated information. Kuel et al. (2016) test several machine-learning approaches in combination with four different sampling methods (No sampling, under-sampling, over-sampling, and synthetic minority over-sampling technique). Their results indicate the lack of practical methods. While they can achieve high precision (i.e. no error rate), this method only uncovers 4% of all needs. Alternatively, their method can achieve fairly better than chance accuracy. Christensen et al. (2017) test different SVM configurations and achieve scores similar to Kuel et al. (2016), with the best model in their study achieving a recall of 0.79 and a precision of 0.42. Overall, further research in machine learning algorithms is required to optimize the process of correctly identifying needs. In return, this will make the massive amount of data available for further processing.

Finally, Timoshenko and Hauser (2019) follow Lee and Bradlow (2011) by analyzing customer reviews. Their research is the most influential work related to this study. Their human-machine hybrid approach aims at extracting specific customer needs from Amazon reviews within single categories. The machine part of their method deploys a CNN to identify need-containing sentences. Then, human experts extract the specific need. The results indicate that within some product categories UGC is at least as valuable as traditional product development methods. The hybrid machine learning approach identifies overall more needs than the expert team, which arguably could be hidden opportunities for companies. One limitation of their study is that for each product category a new model has to be trained, which requires new labeled training data. We plan to close this gap by exploring if certain machine learning models can generalize across product categories, and therefore reduce the costs associated with needmining. For this, we propose the use of alternative natural language processing methodologies, for example, pre-trained transformer models.

2.2 Deep learning based natural language processing

Many different classification techniques and models for textual data exist (e.g neural networks, nearest neighbor classifiers, tree-based methods, naïve Bayes classifiers, support vector machines). SVMs have been proven to be a simple and effective method to achieve strong results (Khan et al., 2010; Kuehl et al., 2016; Christensen et al., 2017; Kühl et al., 2020), and are therefore a popular choice when working on a textual classification problem.

However, in recent years tremendous advancements have been made in the area of NLP (Talmor et al., 2019). This is mainly due to the new transformer architecture first described by Vaswani et al. (2017) and the availability of transfer learning coming to NLP (Liu et al., 2019). Transformers consist of multiple encoders and decoders stacked on top of each other. Encoders process the input sentence and generate a vector representation of the input text. The resulting vector representations are fed into the decoder stack. The outputs of the decoder stack are then used as input by a fully connected neural network, followed by a softmax-layer which makes the final prediction (Vaswani et al., 2017). The Transformer, unlike CNNs or Recurrent Neural Networks, does not rely on sequential processing during the encoding step and therefore lends itself to parallelization (Young et al., 2018). To still capture dependencies in the input that are far away from each other the Transformer uses a modified version of attention, i.e. multi-head attention, which is included in the encoder (Vaswani et al., 2017). Transformer-based architectures were able to break multiple benchmarks in a variety of different NLP tasks (Vaswani et al., 2017; Devlin et al., 2019; Liu et al., 2019) and outperform previous architecture such as CNNs (Young et al., 2018; Hupkes et al., 2020).

Additionally, inductive transfer learning in the form of pre-training brought significant performance gains to many NLP tasks (Liu et al., 2019). In the NLP domain typically Language Modelling is used as the pre-training task (Devlin et al., 2019). For this task, a network, referred to as a Language Model (LM), must predict the masked or next words given a masked or incomplete sentence (Bengio et al., 2003; Devlin et al., 2019). By doing this the model learns rich language features that are encoded into the weights of the neuronal network. The pre-trained LM is then used as a feature extractor to encode the input of other datasets (Hupkes et al., 2020). Next, the encoded input is fed into another neuronal network which does the final task (for example text classification). This pre-training step enables inductive transfer learning (Howard and Ruder, 2018). The model transfers knowledge gathered from solving the previous task (predicting words in sentences) to the final so-called “downstream” task. Interestingly this approach not only enables the model to achieve good results on one type of task but on several different tasks after a small amount of fine-tuning (Radford et al., 2019). This leads to the assumption that these models have greater generalization capability than previous approaches, resulting in models able to predict across product categories with higher performance.

Many different models exist that make use of pre-training and transformers (e.g. Devlin et al., 2019; Liu et al., 2019; Raffel et al., 2020). These models mainly differ in the data used for the pre-training as well as small tweaks in the pre-training objectives, e.g. masked word prediction versus next word prediction. While these models do differ specifics they all show the previously describes feature of being able to achieve good results on serval tasks.

3 Methodology

We collected user reviews for the top 3 most popular products of each major category on Amazon. All data were collected from amazon.com during November 2019. This yielded 432,149 reviews from 115 products spanning 39 categories. We split the reviews into individual sentences using the Natural Language Toolkit in Python (Bird et al., 2009), which resulted in roughly 1,200,000 sentences. After screening the categories, we found that some did not contain any products with typical customer needs (e.g. gift cards or magazine subscriptions). We decided to drop such categories from our dataset. This left us with a total of 25 categories, 75 products, and 315,392 reviews. For the transformer model we used the same tokenization strategy used in the pre-training step. For the other models we tried the same pre-processing steps as Timoshenko and Hauser (2019) which are removing stop words, punctuation, numbers, concatenating frequent combinations of words into n-grams, and dropping the shortest and longest 10% of the sentences. Same as Christensen et al., (2017) we found that removing stop words and concatenating frequent combinations of words reduced the performance of the models. Consequently, we excluded these performance-reducing steps while applying the rest. After pre-processing, the corpus consists of 680,855 sentences.

We manually labeled 2,153 sentences from the Baby-category to identify all sentences in the set that contain a need. The labeling was performed by the authors and student-assistants who were well educated regarding the definition provided earlier in this paper. Each sentence containing a need received a dichotomous “need tag”, where 1 indicated the presence of a need and 0 otherwise. We chose the Baby category because the initial screening revealed a high number of sentences with needs compared to other categories. As the goal of our classification task is to identify needs within sentences, we require a specific set of data that does not suffer from zero inflation of the target variable to prevent the model from favoring the majority class. This produced a balanced dataset with around 41,2% of sentences containing needs. Additionally, we randomly sampled 32 sentences from each category and labeled them the same way. This results in a hold-out set of 800 sentences which we use for the validation out of category. Table 1 displays a subset of the labeled data split by their respective need tag.

Need Containing Sentences	No-Need Containing Sentences
I love that this car seat is lightweight so when it is taken out of the car it is not a struggle.	My son is in love with his new booster seat.
Easy to install and my daughter loves facing forward now!!	my son loves dinosaurs and he didn't know I ordered this booster seat when he saw it he was in love with it
Very easy to assemble plus it isn't very heavy.	I brought 2 of these, one for each vehicle, and I love it.

Table 1. Example selection of the labeled data split by need content.

4 Analysis and Results

As our pre-trained Transformer model we select RoBERTa given its reported performance (Liu et al., 2019). The tested RoBERTa model follows the pretraining and architecture presented by Liu et al. (2019). We compare this model against three baselines, a Bag of Words SVM, a TF-IDF SVM, and a Convolutional Neural Net with Word Embeddings based on Timoshenko and Hauser (2019).

For our analysis, we trained all the presented models on annotated sentences from the baby category. The models were tasked to predict the occurrence of a need in the input sentence, i.e. our binary classification task. We train the models on 2153 annotated sentences using k-fold cross-validation to increase the robustness of the analysis. The reported metrics are the average performance of the k models on their respective test sets. We selected k = 10 as increasing the number of folds did not significantly affect the performance metrics. The models were fitted exclusively on the training set such that the test set can be used to determine how well the model performed in the same product category but also across product categories.

Model	In Category	Out of Category	Literature
RoBERTa	**83%**	**73%**	This study
SVM – BOW	77%	45%	Timoshenko and Hauser (2019), Kuel et al. (2016)
SVM – TFIDF	75%	53%	Timoshenko and Hauser (2019), Kuel et al. (2016)
Word2Vec + CNN	75%	50%	Timoshenko and Hauser (2019)

Table 2. F1 score of the models on the Baby category (In Category) and a combination of 24 product categories (Out of Category).

To compare the model's performance, we use the F1 score (Powers and Ailab, 2011). Choosing this performance measure is a consequence of an unbalanced dataset, as the number of sentences containing customer needs is sparse (Kühl et al., 2020). This makes other measures such as accuracy, which measure the rate of right predictions (true positive and true negative within the entire sample), less meaningful. The F1-score overcomes these limitations as it is defined as the harmonic mean between recall and precision:

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

While the in-category test set only contains sentences within the same category as the training set of the model by definition, the out-of-category test set is evenly composed of all categories other than the training category. Table 2 shows the F1 scores of the model and baselines on both test sets. We observe that the proposed RoBERTa model performs best on all metrics in and out of the category. Interestingly it also has the lowest drop in performance when comparing in and out of the category (10% versus 32%, 22%, and 25%). This supports that the model has the ability to generalize across categories, and indicates the possibility of using a single model for cross-category-needs prediction. Figure 1 displays the F1 scores of the model and the baselines grouped by category. The proposed RoBERTa model performs similarly in related categories such as Camera Photos, Electronics and Computer Accessories, but not generally well. This is to be expected as the consumer needs in the Baby category differs quite largely from some other categories, and with a generally small training size of 2153 sentences. Compared to Timoshenko and Hauser (2019) we receive a relative improvement of 12.2% in F1 score in category.

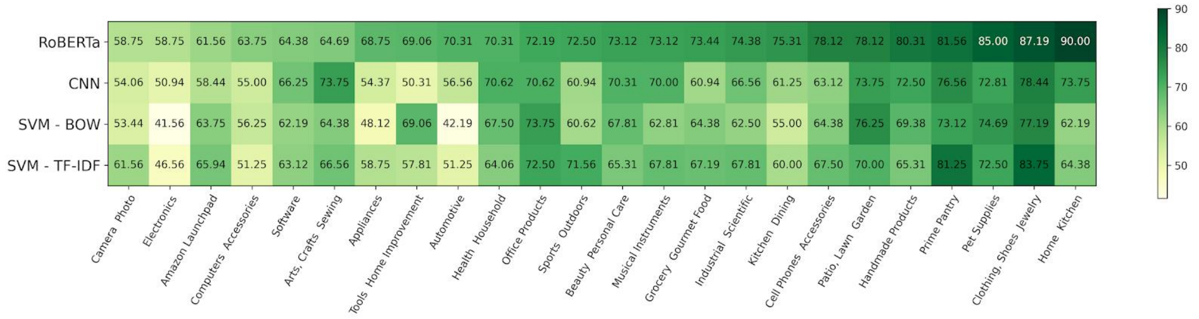

Figure 1. Model and baseline F1 scores on different product categories not occurring in the training set.

The result indicate that RoBERTa should be used as the primary model when employing need identification from UGC, since it outperforms all other models in our evaluation. While RoBERTa performs well on some categories it was not trained on it is still far away from producing good results on all categories (e.g. Electronics). On average there still is a large performance drop when predicting data that is not in the product category of the training data, especially when the target category is very different from the training category. Therefore when the cost of analyzing data through experts is high but labeling data is cheap it can still be advised to label additional data matching the target product category.

The evaluation presented in this study certainly has some limitations. The data source chosen for the models stems from Amazon product reviews, it is unclear if different sources such as Twitter have different characteristics that affect the performance. Additionally, we only evaluate models trained on the Baby product category. Models trained on other product categories could exhibit deviating performances.

In summary, we support the idea of using information systems to analyze user-generated content for consumer needs to make product development more efficient. We build on existing research and introduce pre-trained transformer models into the needmining domain. We further add to efficiency by showing initial signs of generalizability across product categories. Especially the generalizability aspect has not previously been considered in research on digitized customer needs identification and constitutes a theoretical contribution. As a practical contribution, this research can help to reduce the cost of needmining approaches by either enhancing the performance of the model or by reducing the amount of training data needed.

5 Future Research

In the future, we plan to extend the model and train it on more heterogeneous data to further test the limits of the pre-trained approaches in needmining. We plan to proceed in three steps. First, we increase the amount of labeled data for training and evaluation purposes. Timoshenko and Hauser (2019) suggest the use of Amazon's Mechanical Turk. We test the feasibility of their proposal by sourcing a large-scale dataset with MTurk and evaluating its reliability for needs analysis based on literature best practices. To get a diverse and balanced dataset, we plan on annotating all categories equally.

With the extended data, we can proceed with the next step. By increasing and further diversifying the number of training sentences, we expect to generally improve the overall model performance. Additionally, we want to measure the impact of a diversified training data set on cross-category performance. We try to show this by fitting various models on different sets of categories. We also want to consider how human-in-the-loop can be used to ensure that the performance of models doesn’t degrade over time and doesn’t suffer from data drift.

The final step is to extend the evaluation. We plan to cover adversarial attacks such as input reduction (Feng et al., 2018) and Saliency Map Visualization (Sundararajan et al., 2017), as a way to gather deeper insights. While adversarial attacks can be used to “fool” a model by forcing a wrong prediction they also help to understand why a model makes a certain decision. The goal of these methods here is to link words in the sentence to the prediction of the model. These methods can highlight which words in the sentence affected the decision most, increasing our understanding of what factors contribute to the outcome of the model.

6 Conclusion

The goal of this study is to test the performance of pre-trained models, in this case, RoBERTa, in a needmining context and to evaluate the out-of-category performance of all relevant, current models. So far, the empirical analysis shows that RoBERTa outperforms previous models (SVM, CNN) when it comes to both classifying needs within category and out of category. Most remarkably, between the methods, RoBERTa has the lowest drop in performance when applied out of category. Moreover, despite only being trained on data from the baby category, RoBERTa still achieves an F1 score of over 70% on 13 out of 24 of the tested categories. These are highly promising results for our future research, as we already partially show that a generalization across categories is possible.

We believe that diversification of our training data could lead to more accurate out-of-sample predictions, and therefore, generalizability. This both supports and adds to Timoshenko and Hauser’s

(2019) hypothesis that needmining contributes to product development. First, it enhances efficiency, as a single, generalizable training set with the right method is cheaper than building several training sets for single categories. Second, the proposed RoBERTa model achieves higher performance than the previously suggested models, we add to the overall performance of product development by providing a clearer overview of the market.

7 Acknowledgment

We gratefully acknowledge the support of the German Research Foundation (DFG), project FOR 1452 / TP1 SCHO 1321/3-2.

References

Bashir, N., K. N. Papamichail and K. Malik. (2017). "Use of social media applications for supporting new product development processes in multinational corporations." *Technological Forecasting and Social Change*, *120*, 176–183.

Bengio, Y., R. Ducharme, P. Vincent and C. Jauvin. (2003). "A neural probabilistic language model." *Journal of Machine Learning Research*, *3*(Feb), 1137–1155.

Bird, S., E. Loper and E. Klein. (2009). "Natural language processing with python O'reilly media Inc."

Carlsson, S., V. Corvello, M. Inauen and A. Schenker-Wicki. (2011). "The impact of outside-in open innovation on innovation performance." *European Journal of Innovation Management*.

Castellion, G. and S. K. Markham. (2013). "Perspective: New Product Failure Rates: Influence of A rgumentum ad P opulum and Self-Interest." *Journal of Product Innovation Management, 30*(5), 976–979.

Chesbrough, H. (2003). "The logic of open innovation: managing intellectual property." *California Management Review*, *45*(3), 33–58.

Christensen, K., S. Nørskov, L. Frederiksen and J. Scholderer. (2017). "In Search of New Product Ideas: Identifying Ideas in Online Communities by Machine Learning and Text Mining." *Creativity and Innovation Management*, *26*(1), 17–30.

Devlin, J., M.-W. Chang, K. Lee and K. Toutanova. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *ArXiv:1810.04805 [Cs]*.

Feng, S., E. Wallace, A. Grissom II, M. Iyyer, P. Rodriguez and J. Boyd-Graber. (2018). "Pathologies of Neural Models Make Interpretations Difficult." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3719–3728). Brussels, Belgium: Association for Computational Linguistics.

Griffin, A. and J. R. Hauser. (1993). "The voice of the customer." *Marketing Science*, *12*(1), 1–27.

Hoornaert, S., M. Ballings, E. C. Malthouse and D. Van den Poel. (2017). "Identifying new product ideas: waiting for the wisdom of the crowd or screening ideas in real time." *Journal of Product Innovation Management*, *34*(5), 580–597.

Howard, J. and S. Ruder. (2018). "Universal language model fine-tuning for text classification." *ArXiv Preprint ArXiv:1801.06146*.

Hupkes, D., V. Dankers, M. Mul and E. Bruni. (2020). "Compositionality Decomposed: How do Neural Networks Generalise?" *Journal of Artificial Intelligence Research*, *67*, 757–795.

Kannan, P. K. (2017). "Digital marketing: A framework, review and research agenda." *International Journal of Research in Marketing*, *34*(1), 22–45.

Khan, A., B. Baharudin, L. H. Lee and K. Khan. (2010). "A review of machine learning algorithms for text-documents classification." *Journal of Advances in Information Technology*, *1*(1), 4–20.

Krishnan, V. and K. T. Ulrich. (2001). "Product development decisions: A review of the literature." *Management Science*, *47*(1), 1–21.

Kuehl, N., J. Scheurenbrand and G. Satzger. (2016). "NEEDMINING: IDENTIFYING MICRO BLOG DATA CONTAINING CUSTOMER NEEDS." *Research Papers*.

Kühl, N., M. Mühlthaler and M. Goutier. (2020). "Supporting customer-oriented marketing with artificial intelligence: automatically quantifying customer needs from social media." *Electronic Markets*, *30*(2), 351–367.

Lee, T. Y. and E. T. Bradlow. (2011). "Automated marketing research using online customer reviews." *Journal of Marketing Research*, *48*(5), 881–894.

Lilien, G. L., P. D. Morrison, K. Searls, M. Sonnack and E. von Hippel. (2002). "Performance assessment of the lead user idea-generation process for new product development." *Management Science*, *48*(8), 1042–1059.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, … V. Stoyanov. (2019). "Roberta: A robustly optimized bert pretraining approach." *ArXiv Preprint ArXiv:1907.11692*.

Misopoulos, F., M. Mitic, A. Kapoulas and C. Karapiperis. (2014). "Uncovering customer service experiences with Twitter: the case of airline industry." *Management Decision*, *52*(4), 705–723.

Powers, D. and Ailab. (2011). "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation." *Journal of Machine Learning Technologies*, *2*, 37–63.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever. (2019). "Language models are unsupervised multitask learners." *OpenAI Blog*, *1*(8), 9.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, … P. J. Liu. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *Journal of Machine Learning Research*, *21*(140), 1–67.

Roberts, D. L. and M. Candi. (2014). "Leveraging social network sites in new product development: Opportunity or hype?" *Journal of Product Innovation Management*, *31*, 105–117.

Sundararajan, M., A. Taly and Q. Yan. (2017). "Axiomatic attribution for deep networks." In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (pp. 3319–3328). JMLR. org.

Talmor, A., Y. Elazar, Y. Goldberg and J. Berant. (2019). "oLMpics–On what Language Model Pre-training Captures." *ArXiv Preprint ArXiv:1912.13283*.

Timoshenko, A. and J. R. Hauser. (2019). "Identifying customer needs from user-generated content." *Marketing Science*, *38*(1), 1–20.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, … I. Polosukhin. (2017). "Attention is all you need." In: *Advances in neural information processing systems* (pp. 5998–6008).

von Hippel, E. (1986). "Lead Users: A Source of Novel Product Concepts." *Management Science*, *32*(7), 791–805.

Young, T., D. Hazarika, S. Poria and E. Cambria. (2018). "Recent trends in deep learning based natural language processing." *Ieee Computational IntelligenCe Magazine*, *13*(3), 55–75.