Consumer-Oriented Tech Mining: Integrating the Consumer Perspective into Organizational Technology Intelligence—The Case of Autonomous Driving

Marc Egger
Universität zu Köln
egger@wim.uni-koeln.de

Detlef Schoder
Universität zu Köln
schoder@wim.uni-koeln.de

Abstract

To avoid missing technological opportunities and to counteract risks, organizations have to scan and monitor developments in the external environment through a structured process of technology intelligence. Previous approaches in tech mining—the application of text mining for technology intelligence—have primarily focused on the elicitation of technical or legal information from web, patent, or research databases. However, knowledge of consumers' needs, fears, and hopes is a prerequisite for the success of an emerging technology in the marketplace. Thus, we claim that technology intelligence needs to also consider consumers' technology perceptions. Hence, we propose a novel and comprehensive approach to collect user-generated content from the web and apply text mining to derive consumer perceptions. In doing so, we align with an established tech-mining process. This paper illustrates our approach on the emerging technology of autonomous driving and provides an initial indication of concurrent validity.

1. Introduction

Nowadays, companies are facing a situation of shortened product and technology lifecycles and an environment where "disruptive" innovations threaten established business models. As a consequence, companies constantly need to innovate and must also have an eye on technological developments with respect to market opportunities and risks. Still, delivering the latest technology to the market or "technology push" innovations are risky as innovations ultimately have to be accepted by their intended customers. Therefore, many companies' newly offered products and services fail within the first three years after innovation introduction [31]. Finally, both missing technological opportunities and not meeting market needs are major risks to companies' survival as innovations are expensive, failures lead to tremendous sunk costs, and missed technological developments and opportunities might even endanger a company's business model.

To avoid missing such technological opportunities, organizations use so-called "technology intelligence," which means that they scan and monitor technological developments within the external environment. Technology intelligence can be defined as the capture and delivery of technological information as part of a structured process through which an organization develops an awareness of technological threats and opportunities [15]. Typical components of technology intelligence are technology monitoring, competitive technological intelligence (CTI), technology forecasting, road mapping, assessment, and foresight. When text mining is used for technology intelligence, it is denoted as tech mining [6]. Users who benefit from technology intelligence include strategic planners, who have to align emerging technologies with organizational objectives; researchers, inventors, and project managers, who have to be informed and look for possible research partners; intellectual property (IP) mangers, who need information to assess the desirability of patenting disclosure; new product developers and designers, who get help in selecting technological alternatives; and marketing experts, who have to identify new risks and opportunities to leverage products and services [23].

Previous research has proposed various methods and structured processes to elicit technical and legal information from the content and structure of the web, publication records, patents, and research databases. As tech mining is not necessarily restricted to these types of data sources and, in fact, the power of tech mining lies in the combination of text and numerical data to best answer questions that may confront any researcher [6], there is a research opportunity to integrate new data sources and additional perspectives into established technology intelligence processes.

In this article, we argue for integrating information on consumers' technology perceptions from user-generated content (UGC) into technology intelligence. This integration is especially important as knowledge

HICSS

of the potential consumers' needs is a prerequisite for a consumer-oriented product-development process [25]. Successful products stem from the intersection of customer need and technological competencies [23], and strategies are needed to minimize the risk of investing in technologies with low end-user potential [30]. Additionally, innovation draws on knowledge regarding present and future customers and the marketplace. As the consumer is critical to new product development, organizations have to build upon their existing customer base and learn about new customers or about the formerly unrecognized needs of consumers, which is another form of intelligence essential to successful innovation [23]. Due to the risk of creating products that consumers will not accept, it is therefore essential to consider the consumer perspective during all phases of the technology and product lifecycle and communicate insights to technology-development stakeholders. Hence, there is a need for "concurrent design", which refers to forming multifunctional teams to develop new products, processes, and services from the research stage onward, and knowledge of changing technology is vital to such work [23].

With the rise of Web 2.0 and social media, a new information source for consumers' perspective has emerged on the internet. Consumers participate and exchange their ideas, perceptions, and feelings on the web by publishing UGC, which is preserved and recorded digitally and offers an under-exploited information source for detecting new social trends. If UGC is used within technology intelligence, this can be thought of as a modernization of the traditional approach of environmental scanning [3]. This paper contributes to technology and innovation management by proposing a novel and comprehensive approach to integrate consumer perceptions into technology intelligence analyses. The approach applies text mining to UGC, builds upon an established tech-mining process [23], and can be applied short term during all phases of product and technology lifecycle. Using this method, both individuals concerned with emerging technologies and multifunctional teams in organizations can receive a bird's-eye- and easy-to-read view of consumers' technology perceptions.

The remainder of this article is structured as follows: Section 2 provides an overview of related research to underpin our research objective. Section 3 describes our proposed methodology and illustrates both our approach and managerial insights into the emerging technology of autonomous driving. Section 4 then provides an early indication of concurrent validity, and Section 5 closes with a discussion and gives an outlook of future research.

2. Related Research

Overall, previous research can be divided into (2.1) research that considers consumer perceptions but uses traditional "reactive" surveys, (2.2) research that applies automated algorithms on publicly available online data but does not consider the consumer perspective, and (2.3) research that considers online UGC but targets higher-level attributes of consumer behavior than the elicitation of consumers' technology perceptions. Thus, the research gap we bridge with this paper lies in extending the perspective of technology intelligence to the integration of consumer perceptions based on the analysis of online UGC.

2.1. Approaches Considering the Consumer Perspective

The value of integrating consumers' perspectives into technology intelligence efforts with respect to innovation and product development was discussed in several works. Lee and Cho [19] propose a theoretical model for market analysis of convergence products and define a theoretical model for determining consumer perceptions and utility. The model is tested using survey data and conjoint analyses. Using interview data and qualitative analysis through four case studies highlighting the inter-relationships between firms and their communities, Paramentier and Mangematin [22] explore the role of online user communities in co-innovation processes. Hung, Lee, and Wang [12] propose and study a modification of the Delphi method to integrate end-user participation into strategic foresight. They believe that while experts may be well versed in technical knowledge, they tend to neglect consumers' sentiments. These research contributions show the potential of integrating consumers' perceptions within technology intelligence analyses, but in contrast to our proposed approach, they use traditional reactive surveys as the data source.

2.2. Approaches Applying Algorithms on Publicly Available Data

The growth of the web has led to a tremendous increase of new data sources for technology intelligence. The appropriate implementation and use of information technology (IT) tools to (automatically) gather and analyze these data is of key importance for the creation of actionable technology intelligence [30]. Various contributions have used algorithms to derive technology-related information from the structure and content of publicly available (web) data. Using network analysis and bibliometrics, co-citation

networks have been derived from patents and scientific publications [14, 24]. Text mining has been applied on patents to determine the stage of a product's lifecycle [11], to measure the degree of patent novelty [18], and to explore knowledge flows [20]. The content of web search engine results has been investigated using text mining to detect temporal trends and to derive knowledge maps for technology assessment [26]. These contributions underpin the potential and utility of using automated algorithms on publicly available data to inform technology intelligence but refer to publication records, journal articles, or patents and thus focus on the producers' perspective. This opens up the research potential of using text mining to investigate the consumers' perspective by investigating UGC.

2.3. Approaches Utilizing UGC

Few previous contributions integrate consumer-related information from the web into technology intelligence. Jun et al. [13] utilize search engine traffic to analyze technology adoption as traditional life-cycle analysis methods are narrowly focused on the producers' perspective. Albert et al. [1] analyze blog content, proposing a fuzzy logic–based text-mining approach to assess technology maturity. Woo et al. [32] model the diffusion of medical information across forums to identify the major needs of Alzheimer disease patients and their families to support corporate foresight. Although all these contributions show the value of considering the consumer perspective for technology intelligence analyses, they differ from our research in terms of focus and granularity. While we holistically refer to UGC distributed across the web, others focus on specific data sources, such as search engine traffic [13], blog content [1], and forum content [32]. Furthermore technology adoption [13], information diffusion [32], and technology maturity [1] refer to higher-level aspects as a consequence of consumer behavior, whereas we refer to consumer perceptions as a determiner of consumer behavior.

3. Our Proposed Approach to Elicit Consumers' Technology Perceptions

In the following, we propose our approach to elicit consumers' technology perceptions from online UGC. We illustrate our approach on autonomous driving, an emerging technology representing a radical innovation in human transportation.

Porter and Cunningham [23] divide the general process of tech mining into nine steps within the three phases of (1) intelligence, (2) analysis and design, and (3) choice. The intelligence phase involves planning

and collecting the data. Within the analysis and design phase, knowledge is derived from the collected data to solve problems related to technology and/or innovation management. Lastly, the choice phase involves the representation and interpretation of the mined data followed by the use of the results in selecting the right opportunities for the organization [23]. Figure 1 shows the tech-mining process [23] to which we align our approach. The bold font shows the steps the present paper contributes to extend the original process.
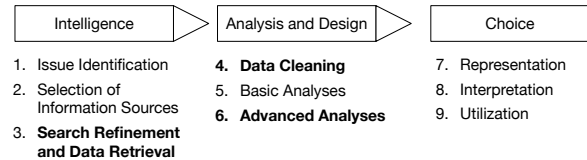

Intelligence	Analysis and Design	Choice
1. Issue Identification	**4. Data Cleaning**	7. Representation
2. Selection of Information Sources	5. Basic Analyses	8. Interpretation
3. **Search Refinement and Data Retrieval**	**6. Advanced Analyses**	9. Utilization

Figure 1. Tech-mining process based on [23]

3.1. Issue Identification

Within the first stage—issue identification—the research object has to be defined. When the research object is not known, an undirected search (denoted as technology/environmental scanning) can usually be performed to elicit candidate research objects [28]. As this is beyond the scope of this article, we assume that the research object has already been identified as an emerging technology based on a radical innovation and "competence-destroying" technological advancements [34].

Within our illustrative study, we consider the emerging technology of autonomous driving as the research object. This seems adequate as the 2015 Gartner Hype Cycle for "Emerging Technologies that organizations should monitor" puts autonomous driving at the top of its "inflated" expectations [35]. Thus, in this context, deriving empirical evidence on consumers' perceptions, such as consumers' needs, wants, fears, and hopes, can be of high value for stakeholders both within the automotive industry and in other industries one would not even think of in the first place. The automotive industry has invested billions of dollars in technology development, but consumer acceptance is vital for the success of autonomous vehicles' market introduction. As such, the automotive industry has to keep a close eye on, for example, consumers' fears to potentially address them before market introduction. However, other industries might also profit from this information to, for instance, create innovative products and services on top of the technology. Here, multifunctional teams of strategic planners, product inventors, designers, and marketing experts would be enabled to identify opportunities to leverage products and services [23].

3.2. Selection of Information Sources

Following the research object definition, information sources have to be selected. For tech mining, information sources usually include patents, journal or conference proceedings, or general HTML-based web sources.

Within the scope of the proposed approach, the information source is online UGC as UGC is the result of consumers' explanation of their personal attitudes and perspectives. UGC can be found on online social networks (OSN), forums, blogs, product review communities, and also any website that enables users to participate by providing comment functionality. Thus, UGC on the web is decentralized as the web itself and often mixes with non-UGC (e.g., editorial content) [7]. Unfortunately, content does not incorporate a directly observable attribute that classifies it as UGC [7], and there is no public search engine equivalent for the automated gathering, querying, retrieving, matching, and displaying of UGC [3]. However, it can be assumed that search engines are the primary entry point for consumers to find relevant links when searching for other consumers' perspectives [3]. Thus, we propose to use general-purpose web search engines (e.g., Google, Yahoo, or Bing) to discover URLs where UGC may also potentially be found and then clean up non-UGC in the data-cleaning stage.

3.3. Search Refinement and Data Retrieval

General-purpose web search engines usually offer a Boolean keyword combination–based query language to retrieve references to documents relating to a research object. It is a cognitive task for the researcher to compose a Boolean query within the search-refinement stage consisting of terms that best characterize target documents and differentiate them from unwanted documents. Afterward, a so-called web crawler has to automatically insert the query into the search engine query input form, trigger the submit button, access the search results list, extract URLs to target documents, and finally retrieve the HTML documents (i.e., websites) directly from the corresponding servers. When search results span multiple results pages, the web crawler also has to utilize the provided pagination function to access the remaining results pages automatically.

For our illustrative study, we developed a web crawler that applies the search query described in Figure 2 to the Google web search engine and collects HTML documents from the URLs that the engine

delivers. Following this approach, we collected 2,129,504 web documents from distinct URLs. Those documents were distributed across 181,922 distinct hosts.

"autonomous driving" OR "self driving car" OR "self driving cars" OR "driverless car" OR "driverless cars" OR "autonomous vehicle" OR "automated driving" OR "piloted driving" OR "driverless car"

Figure 2. Boolean search query

3.4. Data Cleaning

Following data retrieval (3.3), a three-stage data-cleaning process aims to exclude information that is not relevant for the investigation of consumer perceptions.

3.4.1. Cleaning Boilerplate. HTML web documents typically consist of various elements besides the main textual content, which are also known as the "boilerplate" (e.g., site navigation, headers, footers, advertisement elements). As these elements do not contribute valuable information but introduce noise, they have to be detected and cleaned before analysis.

For our study, we apply the boilerplate-detection approach proposed by Kohlschütter [16]. This approach relies on generic properties of HTML documents (e.g., link-to-text ratio) and can be considered quite robust to heterogeneous input documents, which are characteristic of documents collected with our approach (3.3).

3.4.2. Cleaning Duplicates. As websites often provide the same content under different URLs (e.g., print views) or similar content is published on multiple webpages (e.g., press handouts or product descriptions), those duplicates also have to be removed. In agreement with Porter and Cunningham [23], we propose to use "absolute" deduplication by character-wise matching every document with one another. Additionally, we suggest applying approaches that are able to detect "near" duplicates (e.g., [29]).

In our study, we ended up with 536,811 documents after cleaning up boilerplate elements and duplicates as described above.

3.4.3. Cleaning Non-UGC. In Section 3.2, UGC was defined as the data source. Unfortunately, UGC is scattered on the web and often mixes with editorial content even within single web documents (e.g., comments on news articles). When aiming to conduct holistic analyses without having to manually select a few web sources where the presence of UGC can be assumed (e.g., forums), we have to separate UGC from

non-UGC more universally as the volume and heterogeneity of webpages prevent individual consideration. For the German language, Egger et al. [7] show that textual UGC incorporates inherent attributes allowing human annotators to consistently identify UGC and that this task can also be addressed by supervised machine learning.

For our study, we apply this approach on our English language document collection. Therefore, we manually annotated a random sample of 422 documents as being UGC (214) or non-UGC (208) and applied a machine learning classifier. Supporting the results from Egger et al. [7], we received best classification performance using support vector machines (SVM) while performing feature selection on words using document frequency thresholding (DFT) [33] and probability ratios (PRs) [9]. Table 1 summarizes the results of our evaluation using 10-fold cross validation on a linear SVM kernel, which had an accuracy of 92.29% for UGC classification.

Table 1. Classification performance

	UGC	Non-UGC
Precision	97.18%	88.46%
Recall	89.61%	96.84%

Our results also support Egger et al.'s [7] findings but for the English language as UGC can be reliably separated from non-UGC using supervised machine learning. After applying the classifier on our whole document collection, we finally receive 106,305 UGC documents distributed across 31,560 distinct hosts (i.e., websites).

3.5. Basic Analyses

After data cleaning, basic webometric analyses can quantify the UGC concerned with the research object. Quantifying which websites (i.e., sources) consumers' discuss might help marketing and innovation managers learn where consumer communities emerge. Time-series analyses on available documents' publication dates show how the discussion volume has evolved over time or whether seasonal or incidental spikes exist.

In our illustrative, study we observe that the distribution of UGC across sources follows a power law distribution where few sources contribute the bulk of total UGC, and a large number of sources incorporate low quantities of UGC. Overall, 25% (26,929 documents) of the total UGC is found on the top 50 sources while the latter 75% (79,376 documents) is scattered across 31,510 sources. Most of the discussions take place on reddit.com (10,115, 9.5%). Of the rest of the top 50 sources, we find 40%

on general-purpose forums and communities (e.g. fark.com or 4chan.com). Thus, autonomous driving might have already gained popularity in the wider public. About 16% of the UGC occurs on emerging technologies and science sources (e.g., arstechnica.com, tech.slashdot.org, hardforum.com, phys.com, dailytech.com, forums.thegegister.co.uk, gizmag.com, or extremetech.com). UGC from social networks like twitter.com, linkedin.com, facebook.com, or Google Plus amounts to 12%, while content from question-and-answer (Q&A) websites amounts to 10%. Communities concerned with cars and mobility contribute 7% to the top 50 list (e.g., thetruthaboutcars.com, uberpeople.com, autoblog.com, teslamotorsclub.com or transportblog.co.nz). Comments on general news websites, such as forbes.com or economist.com, amount to about 3%. The remainder of the UGC on the top 50 sources is scattered across special interest communities concerned with gaming, finance, and even activism and conspiracy theories.

Time-series analyses reveal that early discussions occurred 10 years ago at the beginning of 2005 when Web 2.0 and social media began to rise. This indicates that our data reflects a comprehensive view of consumers' perspectives. The average number of UGC posts per month rises from 22 in 2005 to 49 at the end of 2009. With the beginning of 2010 (132 UGC posts/month), when Google's driverless car debuted, the amount of UGC posts per month doubles every year leading to a near exponential growth (2,726 UGC posts/month) by the end of 2014. Furthermore, two spikes in UGC volume stand out. The first is in May 2014 when Google revealed its Google car prototype to the public, and the second is in March 2015 when Tesla CEO Elon Musk gave a statement that self-driving cars would someday be outlawed.

3.6. Advanced Analyses

In alignment with the tech-mining process [23], we now describe our core contribution by means of an advanced analysis using text mining to uncover consumers' perception. We assume that consumers' perceptions manifest in words (and phrases) and relationships among them within UGC. These words and phrases can be individuals' names (e.g., Mark Zuckerberg), locations (e.g., Denver, France), companies (e.g., IBM, Apple, Google), institutions (i.e., the University of Scotland), general topics (e.g., price, safety, car), or emotions and actions (e.g., fear, love, buy). In the following, we denote those words and phrases as concepts. As natural language concepts can have direct relationships with each other, we denote those relationships as "driver relation," while

the source of the "driver relation" is referred to as "driver (concept)" (e.g., within the relation of <nice → Car> "nice" is denoted as the "driver (concept)," which has a relationship with the concept "car"). To extract concepts and driver relations from UGC, we propose a three-stage approach of tokenization, part-of-speech (PoS) sequence mining, and sentiment analysis.

3.6.1. Tokenization. First, documents have to be segmented into smaller textual entities, such as sentences and words, using tokenization. Tokenization cuts text into a set of tokens according to specific rules [8]. To receive a set of sentences for each document, we use the tokenization rule of punctuation marks (e.g., exclamation and question marks or periods). As the function of periods for sentence endings is ambiguous (e.g., abbreviations, dates, enumerations, and digits all use periods), we utilize an abbreviation lexicon and hand-crafted rules (to indicate how abbreviations are formed) similar to Grefenstette and Tapanainen [10], who received a sentence boundary recognition rate of 99.07%. Following sentence tokenization, we also tokenize words using whitespace characters (e.g., whitespace, tabulator, or new line). From this, a collection of word-tokenized sentences for each document is received. As in some languages, word entities are composed of multiple subsequent tokens, such as compound nouns (e.g., "steering wheel"), so additional tokenization techniques (e.g. [17]) have to be applied.

3.6.2. PoS Sequence Mining. In this stage, concepts that consumers' are concerned about are extracted from the word-tokenized sentences. Each word token within each sentence of each document gets assigned a certain PoS tag. In our study, we use the decision tree–based PoS-tagging approach proposed by Schmidt [27], which achieves an accuracy of 96.36% on the Penn-Treebank data, is available for various languages, and can easily be extended to additional languages. From each sentence consisting of words, with each word having an assigned PoS tag, we then create a secondary representation solely consisting of PoS tags. Therefore, each sentence receives a secondary representation consisting of a sequence of PoS tags. These PoS-tag-sequences allow us to define rules on a more abstract sentence representation to describe patterns of which words shall be extracted as concepts and drivers. One example is the mapping from the word-tokenized sentence <Tesla, was, visionary> to the PoS-tag sequence of <Noun, Verb, Adjective>. This PoS-tag-sequence describes the pattern to create a rule where the noun (Tesla) represents the concept and the adjective (visionary) represents the driver.

Within our study, we use the patterns and rules described in Figure 3 to extract concepts and related drivers. PoS tags in boxes represent the words that should be extracted as concepts (describing topics in consumers' perception). PoS tags in bubbles represent words that should be extracted as drivers (describing emotions and actions). Patterns in brackets are optional. For readability, we simplified the PoS tags in Figure 3 in contrast to the output from the PoS tagger.
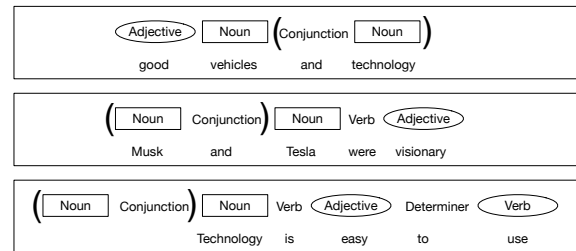
Figure 3. Sequential PoS-patterns

As a result, we end up with 49,176 distinct concepts having at least one driver relation and 1,681 concepts when considering only those concepts that occur in more than 100 documents. Concepts' document frequency describes a Zipf distribution, where few concepts occur frequently but the bulk of concepts occur rarely.

3.6.3. Sentiment Analysis. Determining the emotional orientation (e.g., positive, negative, or neutral) on the concept level is especially promising as this allows one to determine not only which concepts consumers perceive as being important but also the valence of those concepts. Therefore, aspect-oriented sentiment-analysis techniques (synonymously referred to as feature-based or concept-based sentiment analysis) are an integral part of our approach. As sentiment-analysis research rapidly progresses (e.g., in enhancing performance), state-of-the-art aspect-oriented sentiment-analysis techniques might be "plugged in" in our approach. Within our study on autonomous driving, we apply a basic bag-of-words approach. We use a sentiment dictionary such as SenticNet [5] for our extracted driver concepts (usually adjectives and verbs) to transfer a certain sentiment to related concepts (usually nouns or named entities). We also apply basic negation detection by investigating the words surrounding two words while looking for words ending with <n't> or the negation operator <not> for each extracted driver concept (e.g., don't <like> <concept>).

3.6. Representation

As a result of the previous steps, we are left with a set of concepts and drivers for each UGC document. This information is the foundation of various aggregations and representations from which we enumerate a few. The most fundamental aggregation would be to sum up every concept and driver occurrence and provide a visual representation (e.g., table, chart, concept cloud) of the most frequent concepts and their drivers to present the top-of-mind consumer perceptions within the environment of the emerging technology (see Table 2 and Table 3). Besides frequency analyses, concepts' co-occurrence might uncover consumer justification chains (e.g., consumers mentioning "price" (negative) also mention "quality" (positive)). These co-occurrence relationships can be used to construct co-occurrence networks, where network analysis might uncover special concept roles within the consumer perception network (e.g., which concepts act as gatekeepers or bridges or are most central in terms of centrality metrics). As elicited concepts can always be traced back to their source documents, exploring and comparing consumer perceptions by segmentation on various criteria, such as time, concept, emotion, or source, becomes possible. For example, one might compare the perceptions of consumers mentioning Google to those mentioning Audi to explore commonalities and differences in consumers' perceived market structure or for (brand or technology) image analyses.

3.7. Interpretation und Utilization

The utility of investigating consumers' technology perceptions ultimately depends on an adequate interpretation leading to better managerial decisions. Then, the effort of integrating consumers' perceptions into technology intelligence can be justified by the success of an emerging technology in the marketplace or by enhanced competitive advantages. By performing basic analyses on the distribution of UGC across sources, places are uncovered where user interest communities evolve. Public relations and marketing managers learn where to engage with potential customers, which in turn increases marketing efficiency. The distribution of UGC over time allows individuals to investigate the seasonality of consumer interest for effective market-introduction planning or to assess how intensely consumers react to external incidents for adequate public-relationship management. Our proposed advanced analysis reveals which concepts consumers perceive as being important, favorable, or unfavorable. Segmentations allow

explorations from different perspectives and competitive benchmarks. Researchers, engineers, and managers directly or indirectly concerned with an emerging technology are able to get a quick view of consumers' perceptions at any stage of the technology or product lifecycle, including during introduction, growth, maturity, or decline. This knowledge represents diagnostic information on possible consumer behavior throughout the lifecycle phases and helps with the development of strategies targeting to affect the lifecycle. For example, prior to the market-introduction phase, knowledge of consumers' expectations and fears helps prioritize development decisions regarding technology issues and improves design communication strategies for the first technology push. Within the introduction phase, an S-shaped lifecycle curve typically runs flat due to low consumer acceptance and adoption, which often results from unclear technology applications [4]. Information on consumers' needs and wants gives indications of possible applications, which informs strategies to increase the speed of technology adoption. During growth, technology issues might be detected (e.g., when consumers express negative attributes). This detection helps developers adjust technologies on time and prevents the growth from flattening. During the maturity phase, flattening growth might be addressed by considering consumers reports of technology-usage scenarios to design new products and to keep the lifecycle from transitioning to a decline phase.

In our study, we consider autonomous driving, which is at the beginning of its market introduction. Adoption risks due to fundamental concerns or lacking usage scenarios have to be uncovered to ensure successful market introduction. Furthermore, knowledge of consumer perceptions helps prioritize feature development for future technology generations. Tables 2 and 3 show the top 15 most frequent positive and negative concepts and selected driver relations (see Section 3.6.2). Basically, consumers' attitudes toward autonomous driving seem positive as being a "good/great" "thing" and "idea." Also potential "economic" and "societal" benefits are expected from this "technological" and "disruptive" innovation. Consumers perceive "free and valuable time" during driving, which opens up additional market opportunities to fill consumers' free time. Safety might be an important argument for consumer acceptance during market introduction as the driver concept "safe" relates to various other concepts like "car," "vehicle," "driver," "system," or "driving" within the top 15 list. The concept "control" furthermore indicates that keeping "manual" and "full" control is indispensable for many consumers.

Table 2. Top 15 positive concepts

Concept	Driver Relations	# docs
Car	fully autonomous (896), smart (758), safe (706), modern (357), intelligent (158)	7,161
Thing	good/great/best (1,551), interesting (223), right (172), different (168), amazing (107)	3,907
Technology	advanced (353), future (214), innovative (125), amazing (116), revolutionary (69)	3,778
Vehicle	fully autonomous (931), safe (215), intelligent (189), smart (108), efficient (55)	3,373
Way	good/great/best (912), different (258), better (218), easy (175), efficient (141)	3,277
Idea	good/great/best (1,185), interesting (82), innovative (63), futuristic (47)	2,303
Driver	better (312), professional (277), advanced (215), good (183), safe (173), best (74)	2,144
System	autonomous (548), intelligent (103), advanced (83), backup (60), safe (53)	1,793
Benefit	potential (205), great/huge (186), economic (138), societal/social (119), obvious (65)	1,701
Cruise control	adaptive (1,628), advanced (51), intelligent (40), smart (38)	1,699
People	ordinary (161), rich (138), smart (128), willing (113)	1,669
Control	manual (747), full (239), complete (155), total (72), direct (48), immediate (28)	1,665
Time	free (239), good (173), enough (166), right (153), exciting (77), valuable (36)	1,581
Driving	safe (290), efficient (64), free (45), no drunk (38)	1,559
Innovation	technological (327), disruptive (111)	1,446

Table 3. Top 15 negative concepts

Concept	Driver Relations	# docs
Problem	big/huge/major (866), real (200), potential (110)	3,763
Car	less (309), expensive (211), dangerous (87), not ready (42)	2,673
Error	human (2263), responsible (19)	2,359
Issue	legal/regulatory (642), major (135), technical (77), moral (34)	2,327
Driver	bad/terrible (420), drunk (381), blind (77), aggressive (76), inattentive (53), reckless (43), impaired (39)	2,146
People	blind (223), disabled (195), poor (114), few (113), lazy (53), drunk (39)	1,727
Thing	bad (397), stupid (100), dangerous (73)	1,676
Driving	drunk (627), reckless/dangerous (158), bad (81)	1,412
Road	busy (87), rural/narrow (135), dangerous (44)	1,069
Time	(too/so) much (285), wrong (46)	1,064
Traffic	heavy (323), slow (69), high (54), complex (30)	977
Technology	expensive (62), unproven (40), complex (30), not ready (27)	956
Crash	fatal (119), imminent (62), inevitable (44), unavoidable (35), potential (33)	933
Blind spot	tiny (8), weak (7), dangerous (5)	911
Idea	bad (228), crazy (80), stupid (45)	895

Negative concepts (Table 3) also show consumers' concerns regarding fundamental and potential safety problems (i.e., danger) and errors due to system and environmental complexity. Nevertheless, concepts carrying a negative valance also point to technology advantages and opportunities, such as "human errors" and "drunk," "reckless," and "bad" driving being prevented. Also "blind" and "disabled" persons are perceived as being consumer groups profiting from this new technology. Traffic efficiency is expected to increase, leading to "fewer (lesser) cars" on the road. Nevertheless, consumers do not expect an early market introduction of fully autonomous cars due to "legal" and "regulatory" issues and expect "too much time" until market introduction. They also expect the technology to be "too expensive," thereby excluding "poor people" from this technology.

Based on these findings, possible managerial implications for the market-introduction phase could be the following: Invest in the development of safety mechanisms for future technology generations, and take advantage of the economic potential from consumers' free time. Build trust in the technology by clarifying how autonomous systems deal with the complexity of their environment and how they solve accident situations reasonably. Leverage perceived use cases (e.g., no need to worry about drunk driving) and enable new customer groups (e.g., disabled or blind people). Think about new consumption models (e.g., share economy) as the price of the technology might be a limiting attribute preventing wider adoption.

4. Comparison Study

To determine whether the proposed approach might act as an analytical and diagnostic tool for eliciting consumers' technology perceptions, we conducted a survey-based comparison study. This study was performed to compare the results of an established consumer perception–elicitation instrument with the results of our proposed approach. With respect to validation, this gives a first hint at concurrent validity by comparing a validated measurement instrument to a measurement instrument that will be tested using a common measurement variable [2]. The measurement instrument to be tested is our proposed approach. For the validated measurement instrument, we use "free elicitation" where respondents are free to express any answer that comes their mind when presented with a certain stimulus [21]. As the stimulus, we presented participants with the following question: "Which terms/concepts initially come to your mind when you think of autonomous driving?" As the measurement variable, we define consumers' technology perceptions based on expressed words and phrases (denoted as "concepts" and "drivers"). As the indicator for concurrent validity, we use the amount of overlapping concepts and drivers elicited from the two instruments. We recruited 41 German graduate students between the ages 19 and 38 years (Mean: 23.38) for the free elicitation study. In response to our stimulus, participants reported 92 concepts, where we aggregated synonymous concepts (e.g., "lack of control" and "no control") to arrive at 67 concepts. The top 10 stated concepts were "future" (11), "advancement" (6), "safety" (5), "Google" (5), "Tesla" (5), "self-driving"

(5), "autonomously" (5), "car" (5), "technology" (3), and "artificial intelligence" (3). When determining the concept overlap, we included those concepts that were mentioned by at least two participants (23). As a result, we observe that 80% of the concepts from free elicitation were found among the top 50 concepts elicited with our proposed approach (100% within the top 1000 out of 49,176 concepts). Thus, our approach and the established measurement instrument (free elicitation) lead to similar results, which seems to give an early indication of concurrent validity.

5. Discussion and Conclusion

In this article, we proposed a novel approach to integrate consumers' perceptions of emerging technologies into technology intelligence. In doing so, we aligned our approach with an established tech-mining process. We illustrated the utility of our approach on the emerging end-user technology of autonomous driving and gave a first hint at concurrent validity by conducting a comparison study using a survey-based consumer perception–elicitation instrument. Our initial results indicate a large overlap with the results of the established instrument. Additionally, our approach opens up the new data source of UGC to consumer-perception elicitation for technology and, thus, contributes to a structured integration of the consumer perspective into technology intelligence. The approach has strength as it is effective, can be applied without the need to recruit or interview consumers, can be performed in a short time, and allows for the analysis of a large quantity of data due to automation. Furthermore, the availability of metadata from UGC (e.g., the publication date) allows researchers to slice and dice in real time, which opens up additional applications. Nevertheless, limitations exist as the approach highly depends on the availability of public UGC. It is only applicable if consumers already discuss an emerging technology on the web. Unfortunately, this generally only occurs for technologies that might directly affect end users or are already in a later stage of their lifecycle. Furthermore, the validity of the comparison study might be limited as we conducted the study on German-speaking participants while our approach was applied on English-language UGC. Therefore, although early indications of concurrent validity are promising, more in-depth validation studies incorporating participants of similar demographics and the same native language are crucial for further research. These studies might also not only consider the overlap of elicited concepts but also compare the strength, favorability, or ranking of concepts. Finally, we hope that our research encourages others to contribute to the topic of structurally integrating the consumer perspective into technology intelligence analyses to finally allow organizations to have an eye on the consumer to reduce the risk of investing in technologies with low end-user potential and neglecting consumers' needs, wants, hopes, and fears.

10. References

[1] Albert, T., Moehrle, M.G., and Meyer, S. Technology maturity assessment based on blog analysis. *Technological Forecasting and Social Change 92*, (2015), 196–209.

[2] American Psychological Association, *Technical recommendations for psychological tests and diagnostic techniques*. Part 2 Psychological Bulletin, 1954.

[3] Cachia, R., Compañó, R., and Da Costa, O. Grasping the potential of online social networks for foresight. *Technological Forecasting and Social Change 74*, 8 (2007), 1179–1203.

[4] Callon, M. The state and technical innovation: a case study of the electrical vehicle in France. *Research Policy 9*, 4 (1980), 358–376.

[5] Cambria, E., Olsher, D., and Rajagopal, D. SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. *Twenty-eighth AAAI conference on artificial intelligence,* (2014).

[6] Cunningham, S.W., Porter, A.L., and Newman, N.C. Special issue on tech mining. *Technological Forecasting and Social Change 73*, 8 (2006), 915–922.

[7] Egger, M., Lang, A., and Schoder, D. Who Are We Listening to? Detecting User-generated Content (UGC) on the *Web. 23th ECIS 2015, Münster,* (2015).

[8] Feldman, R. and Sanger, J. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, 2006.

[9] Forman, G. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *J. Mach. Learn. Res. 3*, (2003), 1289–1305.

[10] Grefenstette, G. and Tapanainen, P. What is a word, What is a sentence? Problems of Tokenization. *COMPLEX, Budapest, Hungary,* (1994), 7–10.

[11] Haupt, R., Kloyer, M., and Lange, M. Patent indicators for the technology life cycle development. *Research Policy 36*, 3 (2007), 387–398.

[12] Hung, C.-Y., Lee, W.-Y., and Wang, D.-S. Strategic foresight using a modified Delphi with end-user participation: A case study of the iPad"s impact on Taiwan"s

PC ecosystem. *Technological Forecasting and Social Change 80*, 3 (2013), 485–497.

[13] Jun, S.-P., Yeom, J., and Son, J.-K. A study of the method using search traffic to analyze new technology adoption. *Technological Forecasting and Social Change 81 IS* , (2013), 82–95.

[14] Kay, L., Newman, N., Youtie, J., Porter, A.L., and Rafols, I. Patent overlay mapping: Visualizing technological distance. *Journal of the Association for Information Science and Technology 65*, 12 (2014), 2432–2443.

[15] Kerr, C.I.V., Mortara, L., Phaal, R., and Probert, D.R. A conceptual model for technology intelligence. *International Journal of Technology Intelligence and Planning 2*, 1 (2006), 73.

[16] Kohlschütter, C., Fankhauser, P., and Nejdl, W. Boilerplate detection using shallow text features. ACM (2010), 441–450.

[17] Lapata, M. and Lascarides, A. Detecting Novel Compounds: The Role of Distributional Evidence. *Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1, Stroudsburg, PA, USA,* (2003), 235–242.

[18] Lee, C., Kang, B., and Shin, J. Novelty-focused patent mapping for technology opportunity analysis. *Technological Forecasting and Social Change 90, Part B*, (2015), 355–365.

[19] Lee, M. and Cho, Y. Consumer perception of a new convergence product: A theoretical and empirical approach. *Technological Forecasting and Social Change 92 IS* , (2015), 312–321.

[20] No, H.J., An, Y., and Park, Y. A structured approach to explore knowledge flows through technology-based business methods by integrating patent citation analysis and text mining. *Technological Forecasting and Social Change 97 IS* , (2015), 181–192.

[21] Olson, J.C. and Muderrisoglu, A. The Stability of Responses Obtained By Free Elicitation: Implications For Measuring Attribute Salience and Memory Structure. *Advances in Consumer Research 6*, (1979), 269–275.

[22] Parmentier, G. and Mangematin, V. Orchestrating innovation with user communities in the creative industries. *Technological Forecasting and Social Change 83 IS* , (2013), 40–53.

[23] Porter, A.L. and Cunningham, S.W. *Tech Mining: Exploiting New Technologies for Competitive Advantage*. John Wiley & Sons, 2004.

[24] Rafols, I., Porter, A.L., and Leydesdorff, L. Science overlay maps: A new tool for research policy and library management. *... for information Science and ...*, (2010).

[25] Roland Ortt, J., Langley, D.J., and Pals, N. Exploring the market for breakthrough technologies. *Technological Forecasting and Social Change 74*, 9 (2007), 1788–1804.

[26] Sasson, E., Ravid, G., and Pliskin, N. Text Mining and Temporal Trend Detection on the Internet for Technology Assessment: Model and Tool. *22st ECIS 2014, Tel Aviv, Israel,* (2014).

[27] Schmid, H. Probabilistic part-of-speech tagging using decision trees. *international conference on new methods in language processing,* (1994), 44–49.

[28] Schuh, G. and Klappert, S., eds. *Technologiemanagement*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[29] Theobald, M., Siddharth, J., and Paepcke, A. SpotSigs: robust and efficient near duplicate detection in large web collections. ACM (2008), 563–570.

[30] Veugelers, M., Bury, J., and Viaene, S. Linking technology intelligence to open innovation. *Technological Forecasting and Social Change 77*, 2 (2010), 335–343.

[31] Wilke, J. and Sorvillo, N. Targeting early adopters--a means for new product survival. *ACNielsen BASES publication*, (2005).

[32] Woo, J., Lee, M.J., Ku, Y., and Chen, H. Modeling the dynamics of medical information through web forums in medical industry. *Technological Forecasting and Social Change 97 IS* , (2013), 77–90.

[33] Yang, Y. and Pedersen, J.O. A comparative study on feature selection in text categorization. *ICML* (1997), 412–420.

[34] Yeo, W., Kim, S., Park, H., and Kang, J. A bibliometric method for measuring the degree of technological innovation. *Technological Forecasting and Social Change 95 IS* ,, 152–162.

[35] Gartner's 2015 Hype Cycle for Emerging Technologies Identifies the Computing Innovations That Organizations Should Monitor. *gartner.com*. http://www.gartner.com/newsroom/id/3114217.