MASTER

Identifying lead users and their insights on social media

case study from the smartphone domain

Zika, A.

Award date:
2016

Link to publication
Identifying Lead Users and their Insights on Social Media

Case Study from the Smartphone Domain

in partial fulfillment of the requirements for the degree of
Master of Science in Innovation Management

Author: Adam Zíka (zika.adam@gmail.com), 0925781, TU/e
Supervisors: dr. Sarah Gelper, Department of Industrial Engineering & Innovation Sciences, TU/e
Prof. dr. Mykola Pechenizkiy, Department of Computer Science, TU/e
dr. Jorn Bakker, Department of Data Science, Coosto

25. 10. 2016 in Eindhoven
Acknowledgments

This project could not have been performed with a lot of dedication and effort. I am very grateful to all who supported me. My special thanks belongs to my family, who has always stood behind me. I would like to also express gratitude to Sarah Gelper, Mykola Pechenizkiy and Jorn Bakker for an amazing technical as well as personal guidance. I would like thank to Sarah especially for her open-minded and cordial supervision. To Mykola and Jorn, I would like to thank notably for their courage and patience of supervising me, a non-computer science student. I would like to also thank to my friends for their help and encouragement, especially to Roel Kerkhofs, Sanjin Pajo, Pavel Jungman, Maaike Schuitema, Rens Melsen, Martin Kantor, and Roman Schejbal. Finally, I would like to thank to Coosto and all great people there.
Abstract

Traditional lead user identification techniques are not suitable to be employed to search for lead users on social media. To address this gap, this research inspects which metrics can help with lead user identification when Big Data is involved. The research is performed as a case study from the smartphone domain based on data from a product-oriented community forum about iPhones. The proposed metrics include metrics which have and have not been previously analyzed in relationship with lead userness. The previously analyzed metrics are outdegree centrality, indegree centrality, betweenness centrality and latent-attributes. Latent-attributes metric aims to capture the degree of user’s unfulfilled needs about missing features of a product (i.e., latent attributes) experienced prior to other people. One existing and two new approaches how to extract latent attributes from unstructured text are introduced. The existing approach is based on template matching and it is not able to extract any latent attribute. The new approaches use word embeddings of Word2vec and they enable extraction of 19 latent attributes in a short time. The metrics which have not be analyzed before are domain-interest, closeness centrality, hub, authority, pagerank, votes and specification-attributes-in-time. Domain-interest and specification-attributes-in-time are newly defined metrics. Domain-interest refers to the proportion of messages about a product category of interest. Specification-attributes-in-time scores users based on a weighted frequency count of specification attributes in users’ messages. The newly suggested attribute weighting turns out not to be useful for lead user identification. With respect to the main research focus, results show that the top 100 scoring users on indegree centrality, hub, pagerank, authority, and betweenness centrality metrics contain at least 10 lead users, a satisfactory amount for a typical project with lead user involvement. Additionally, votes, closeness centrality, outdegree centrality and specification-attributes-in-time metrics contain at least 10 lead users among 150 the top scoring users. Latent-attributes and domain-interest metrics are less useful. The correlation analysis does not reveal any significant relationship between any metric and lead userness. Finally, Lead User Search Assistant (LUSA), a tool to help domain experts to manually assess the top scoring users in terms of their lead userness is built. It is proposed it can reduce lead user identification procedure to hours.

Keywords: Lead users, automated identification, user innovation, social network analysis, latent feature extraction, word embeddings.
Executive Summary

Consumer innovation is a pervasive business phenomenon [62]. In the United Kingdom for instance, consumers are major sources of innovation [62]. British consumers devote 44% more resources than all of the commercial entities are spending on R&D of consumer products [62]. One of the categories of user innovators are lead users. Formally, lead users are users whose “present strong needs will become general in a marketplace months or years in the future” and who “significantly benefit from obtaining the solution to their needs” [61].

There are various benefits of lead user involvement into new product development. Lead users can help a company for instance to develop 30% more innovative solutions [60], shorten the development time by 44% and cost by 49% [24], or reach eight times higher sales and two times higher market share [37]. To take advantage of these benefits, lead users need to be identified. That is not an easy task. In fact, lead user identification is the most crucial issue and time consuming phase of lead user involvement [7, 8, 53]. It is argued that this is because of two reasons. First, because these innovative users are rare [63]. Second, because the traditional identification methods involve a large amount of a manual effort, because they make use of user surveys and expert analysis [8, 58, 60, 65]. This raises a question how lead user identification can be done in less time and lower cost.

The amount of data on the Internet has been growing rapidly. There are 2.5 quintillion bytes generated every day; 90% of which has been produced in last two years [27]. With this amount of data, it has become feasible to search for lead users on the Internet, especially in specialized product-oriented online communities [16]. However, the big amount of information also stresses a need for emergence of (semi-)automated identification techniques. In the context of a product-oriented community forum, such (semi-)automated technique can be established for instance by using metrics. Metrics represent scores assigned to every user. If users could be filtered or even ranked with respect to their lead userness (i.e., the degree to which a user is a lead user) according to these scores, using metrics would substantially help with lead user identification.

Research Questions

The main research question and sub-questions are stated as follow.

*Which metrics can help to identify lead users in the context of a product-oriented community forum?*

1. *Which metrics contain a sufficient amount of lead users among the top scoring users of these metrics?*

2. *Which metrics are positively related with lead userness?*

With regards to the first sub-question, a sufficient amount of lead users is determined based on a typical amount of involved lead users into new product development projects. This is usually between 8 to 10 lead users [64]. In this research, the upper boundary of 10 lead users is recognized as the sufficient amount.
The second sub-question can be interpreted in a way that if a metric is related with lead userness, then the top scoring users contain in average more lead users than lower ranked user groups.

Both sub-questions are mutually exclusive. If a metric contains a certain amount of lead users among the top scoring users, that does not mean it is positively related with lead userness. The same works the other way around.

**Methodology**

The methodology involves several steps to inspect which metrics can help to identify lead users. First, in a theoretical perspective, it is proposed, which metrics might be suitable for lead user identification and how to assemble them. The metrics are hypothesized to capture some of the characteristics of lead users. Hence, lead users are expected to score higher on the analyzed metrics, because they possess these characteristics in greater degree than ordinary users. Inspecting various metrics in relation with lead userness is beneficial, because different characteristics of lead users can be captured and thus different subgroups of lead users can be potentially identified. The metrics are presented in three groups: network, text and attribute metrics.

The network metrics are built from discussions between community users. They express users’ position in network from different perspectives. The network metrics include outdegree centrality, indegree centrality, domain-interest, betweenness centrality, closeness centrality, hub, authority, and pagerank. Domain-interest is a newly proposed metric and it refers to the proportion of messages about a product category of interest. The other metrics are common network metrics. Some are simple like for instance outdegree centrality, which expresses an amount of messages posted by a user in a forum. Others are more complicated like for instance pagerank, which expresses the degree of a user’s expertness.

The text metrics intend to capture lead userness via a prevalence of product attributes in unstructured text. The analyzed text metrics are latent-attributes and specification-attributes-in-time. Latent-attributes metric aims to capture one of the defining characteristics of lead users, the degree of user’s unfulfilled needs about missing features of a product (i.e., latent attributes) experienced prior to other people [59]. The more a user talks about latent attributes, the more he or she is considered a lead user. The specification of latent attributes is crucial. Therefore, one existing and two new approaches how to extract latent attributes from text are introduced. The existing approach is based on template matching. The new approaches use word embeddings of Word2vec. Specification-attributes-in-time scores users based on a weighted frequency count of specification attributes (i.e., product specification attributes of existing products) in users’ messages. Reflecting the model of innovation diffusion, in which the earlier adopters are considered more innovative [51], attributes posted earlier relatively to their release dates have a higher weight. To inspect whether the proposed weighting function is useful for lead user identification, this metric is compared with a baseline, which is just like specification-attributes-in-time, but omits the attribute weighting.

The attribute metrics are metrics derived from meta-data of messages. The attributes metrics include only one metric, votes. Votes metric represents an aggregated count of a user’s votes.

Second, in a case study, the metrics are computed and ground-truth labels of lead user-
ness are obtained. The labels are obtained via screening, a traditional lead user identification method. It is inspected whether the top scoring users of each metric contain the sufficient amount of lead users and also whether the metrics are positively related with lead userness.

Finally, Lead User Search Assistant (LUSA) is designed. It is a tool to help domain experts with lead user identification.

Case Study & Conclusion

The case study is performed in the domain of smartphones. Data originates from the social media site Reddit, specifically from a thread about iPhones\(^1\).

To first comment on intermediate research results, the extraction of latent attributes revealed that the previously used technique based on template matching is not able to extract any attribute. In contrast, the two newly proposed approaches based on Word2vec enabled extraction of 19 latent attributes. This can be interpreted that 19 new ideas how to expand an iPhone are provided. These ideas represent needs which certain community users possess prior to other people.

Specification-attributes-in-time was compared with a baseline. It appeared that specification-attributes-in-time does not perform better the baseline and therefore the attribute weighting is not useful for lead user identification. Despite that, specification-attributes-in-time approaches lead user identification from a completely different perspective. It is perceived that it has a potential to complement other metrics and identify a different sub-group of lead users.

With respect to the main research focus, results showed that the top 100 scoring users on indegree centrality, hub, pagerank, authority, and betweenness centrality metrics contain at least 10 lead users, a satisfactory amount for a regular project with lead user involvement \([64]\). That means that these metrics can help with lead user identification in a way that the size of population to be examined can be lowered from potentially all 113,030 to 100 community users. In practice, obviously not all community users are inspected. In the traditional identification method screening, approximately 2000 community users need to be approached to identify around 20 lead users \([8, 38]\). The chance of sampling a lead user is therefore around \(\frac{1}{100}\). That is 10 times less comparing to a chance of \(\frac{1}{10}\) when the top 100 scoring users are examined. Additionally, votes, closeness centrality, outdegree centrality and specification-attributes-in-time metrics contain at least 10 lead users among 150 the top scoring users and are considered to boost lead user identification process as well. Latent-attributes and domain-interest metrics turned out to be less useful. The correlation analysis did not reveal any significant relationship between any of the metrics and lead userness. That means that a group of lower ranked users might contain more lead users than a group of the top scoring users. This outcome might be explained by the limited data size or the sampling strategy used to obtain labels.

Lastly, a tool LUSA to help domain experts with lead users identification was built. LUSA facilitates latent attribute identification. It also enables to rank users based on a selected metric and then it provides an interface for domain experts to manually assess the top scoring users in terms of their lead userness. The main advantage of LUSA is that it analyzes all community users and presents ones with a higher potential to be lead users. It is expected that a domain

\(^1\)http://www.reddit.com/r/iphone
expert can shorten lead user identification time from a usual duration of 8.5 man-days \[8\] to hours.

This project has been done as a case study of smartphones. It is therefore questionable whether the results would hold in different domains. It is expected that it would be more difficult to establish this study in a low tech and low involvement product domain. This is because such domains are less vividly discussed on social media, so less data would be available for the analysis.
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgments</td>
<td>i</td>
</tr>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Executive Summary</td>
<td>iii</td>
</tr>
<tr>
<td><strong>1 Introduction</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1 Methodology</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Case Study</td>
<td>3</td>
</tr>
<tr>
<td>1.3 Structure</td>
<td>4</td>
</tr>
<tr>
<td><strong>2 Literature Review</strong></td>
<td>4</td>
</tr>
<tr>
<td>2.1 Lead Users &amp; New Product Adoption</td>
<td>5</td>
</tr>
<tr>
<td>2.2 Lead User Involvement</td>
<td>6</td>
</tr>
<tr>
<td>2.3 Lead User Identification</td>
<td>9</td>
</tr>
<tr>
<td><strong>3 Metrics</strong></td>
<td>14</td>
</tr>
<tr>
<td>3.1 Network Metrics</td>
<td>14</td>
</tr>
<tr>
<td>3.2 Text Metrics</td>
<td>18</td>
</tr>
<tr>
<td>3.3 Attribute Metrics</td>
<td>23</td>
</tr>
<tr>
<td><strong>4 Case Study</strong></td>
<td>24</td>
</tr>
<tr>
<td>4.1 Domain &amp; Data</td>
<td>24</td>
</tr>
<tr>
<td>4.2 Computation of Metrics</td>
<td>25</td>
</tr>
<tr>
<td>4.3 Obtaining Labels</td>
<td>29</td>
</tr>
<tr>
<td>4.4 Results &amp; Discussion</td>
<td>31</td>
</tr>
<tr>
<td>4.5 Lead User Search Assistant</td>
<td>33</td>
</tr>
<tr>
<td><strong>5 Conclusion</strong></td>
<td>36</td>
</tr>
<tr>
<td>5.1 Main Contributions</td>
<td>38</td>
</tr>
<tr>
<td>5.2 Limitations &amp; Future Research</td>
<td>39</td>
</tr>
<tr>
<td><strong>References</strong></td>
<td>40</td>
</tr>
<tr>
<td><strong>Appendices</strong></td>
<td>45</td>
</tr>
<tr>
<td>A Lead User Characteristics</td>
<td>45</td>
</tr>
<tr>
<td>B Literature Review about Relation Extraction</td>
<td>49</td>
</tr>
<tr>
<td>B.1 Supervised Learning Approach</td>
<td>49</td>
</tr>
<tr>
<td>B.2 Semi-supervised Learning Approach</td>
<td>49</td>
</tr>
<tr>
<td>B.3 Unsupervised Learning Approach</td>
<td>50</td>
</tr>
</tbody>
</table>
1 Introduction

Consumer innovation is a pervasive business phenomenon [62]. In the United Kingdom for instance, consumers are major sources of innovation [62]. British consumers devote 44% more resources than all of the commercial entities are spending on R&D of consumer products [62]. One of the categories of user innovators are lead users. Formally, lead users are users whose “present strong needs will become general in a marketplace months or years in the future” and who “significantly benefit from obtaining the solution to their needs” [61].

There are various benefits of lead user involvement into new product development. Lead users can help a company for instance to develop 30% more innovative solutions [60], shorten the development time by 44% and cost by 49% [24], or reach eight times higher sales and two times higher market share [37]. To take advantage of these benefits, lead users need to be identified. That is not an easy task. In fact, lead user identification is the most crucial issue and time consuming phase of lead user involvement [7, 8, 53]. It is argued that this is because of two reasons. First, because these innovative users are rare [63]. Second, because the traditional identification methods (i.e., pyramiding, screening, signalizing, netnography, crowdsourcing) involve a large amount of a manual effort, because they make use of user surveys and expert analysis [8, 58, 60, 65]. This raises a question how lead user identification can be done in less time and lower cost.

The amount of data on the Internet has been growing rapidly. There are 2.5 quintillion bytes generated every day; 90% of which has been produced in last two years [27]. With this amount of data, it has become feasible to search for lead users on the Internet, especially in specialized product-oriented online communities [16]. However, the big amount of information also stresses a need for emergence of (semi-)automated identification techniques. In the context of a product-oriented community forum, such (semi-)automated technique can be established for instance by using metrics. Metrics represent scores assigned to every user. If users could be filtered or even ranked with respect to their lead userness (i.e., the degree to which a user is a lead user) according to these scores, using metrics would substantially help with lead user identification.

Research Questions

The main research question and sub-questions are stated as follow.

Which metrics can help to identify lead users in the context of a product-oriented community forum?

1. Which metrics contain a sufficient amount of lead users among the top scoring users of these metrics?

2. Which metrics are positively related with lead userness?

With regards to the first sub-question, a sufficient amount of lead users is determined based on a typical amount of involved lead users into new product development projects. This is usually between 8 to 10 lead users [64]. In this research, the upper boundary of 10 lead users is recognized as the sufficient amount.
The second sub-question can be interpreted in a way that if a metric is related with lead userness, then the top scoring users contain in average more lead users than lower ranked user groups.

Both sub-questions are mutually exclusive. If a metric contains a certain amount of lead users among the top scoring users, that does not mean it is positively related with lead userness. The same works the other way around.

1.1 Methodology

The methodology involves several steps to inspect which metrics can help to identify lead users (see Figure 1). First, in a theoretical perspective, it is proposed, which metrics might be suitable for lead user identification and how to assemble them. The metrics are hypothesized to capture some of the characteristics of lead users. Hence, lead users are expected to score higher on the analyzed metrics, because they possess these characteristics in greater degree than ordinary users. Inspecting various metrics in relation with lead userness is beneficial, because different characteristics of lead users can be captured and thus different subgroups of lead users can be potentially identified. The metrics are presented in three groups: network, text and attribute metrics.

The network metrics are built from discussions between community users. They express users’ position in network from different perspectives. The text metrics intend to capture lead userness via a prevalence of product attributes in text. The attribute metrics are metrics derived from meta-data of messages. There are 11 metrics examined in total.

Figure 1: Research Methodology.

Indegree centrality, outdegree centrality, betweenness centrality, and latent-attributes metrics have been examined in relation with lead userness in previous research. The first three listed metrics are common network metrics. Latent-attributes metric aims to capture the degree of user’s unfulfilled needs about missing features of a product (i.e., latent attributes) experienced prior to other people [59]. The more a user talks about latent attributes, the more he or she is considered a lead user. The specification of latent attributes is crucial. Therefore, one existing and two new approaches how to extract latent attributes from unstructured text are introduced. The existing approach is based on template matching. The new approaches use word embeddings of Word2vec.
The research also investigates metrics which have not been examined before. They are domain-interest, closeness centrality, hub, authority, pagerank, specification-attributes-in-time, and votes. Domain-interest and specification-attributes-in-time are newly defined metrics. Domain-interest refers to the proportion of messages about a product category of interest. Specification-attributes-in-time scores users based on a weighted frequency count of specification attributes (i.e., product specification attributes of existing products) in users’ messages. Reflecting the model of innovation diffusion, in which the earlier adopters are considered more innovative [51], attributes posted earlier relatively to their release dates have a higher weight. To inspect whether the proposed attribute weighting is useful for lead user identification, this metric is compared with a baseline, which is like specification-attributes-in-time, but omits the weighting. Votes metric represents an aggregated count of a user’s votes. The other metrics are common metrics from social network analysis.

Second, in a case study, the metrics are computed and ground-truth labels of lead user-ness are obtained. The labels are obtained via screening, a traditional lead user identification method. It is inspected whether the top scoring users of each metric contain the sufficient amount of lead users and also whether the metrics are positively related with lead userness.

Finally, Lead User Search Assistant (LUSA) is designed. It is a tool to help domain experts to identify lead users and also latent attributes.

1.2 Case Study

The case study is performed in the domain of smartphones. Data originates from the social media site Reddit, specifically from a thread about iPhones[^1].

The extraction of latent attributes revealed that the previously used technique based on template matching is not able to extract any attribute. In contrast, the two newly proposed approaches based on Word2vec enabled extraction of 19 latent attributes in a very short time.

Specification-attributes-in-time was compared with a baseline, which omits the weighting function. It appeared that specification-attributes-in-time metric does not perform better than the baseline and therefore the attribute weighting is not useful for lead user identification.

With respect to the main research focus, results showed that the top 100 scoring users on indegree centrality, hub, pagerank, authority, and betweenness centrality metrics contain at least 10 lead users, a satisfactory amount for a regular project with lead user involvement [64]. These metrics can therefore reduce the size of population to be analyzed by domain experts to a manageable sample of 100 community users. Additionally, votes, closeness centrality, outdegree centrality and specification-attributes-in-time metrics contain at least 10 lead users among 150 the top scoring users and can help to identify lead users as well. Latent-attributes and domain-interest metrics turned out to be less useful. The correlation analysis did not reveal any significant relationship between any of the metrics and lead userness. That means that a group of lower ranked users might contain more lead users than a group of the top scoring users.

A tool LUSA was built. It is suggested that it can reduce lead user identification time, which usually takes around 8.5 man-days [65], to hours.

[^1]: [http://www.reddit.com/r/iphone](http://www.reddit.com/r/iphone)
1.3 Structure

The thesis is structured as follows. A literature review about lead users is presented in Section 2. Metrics are theoretically defined in Section 3. A case study is introduced in Section 4. It includes information about computation of the metrics and obtaining labels. Furthermore, results and an explanation of LUSA are presented. In Section 5, conclusions, limitations and future research directions are provided.

2 Literature Review

Consumer innovation is an important aspect of New Product Development (NPD) [62]. As mentioned earlier, in the United Kingdom for instance, consumers are the main source of innovation [62]. Next to that, Lüthje [39] found that 22% of surgeons develop new products for an in-house use. In the area of extreme sporting communities, consumers are inventors as well [15]. They participate in the development of 32% of products. For instance, the first versions of snowboards, skateboards, and windsurfboards were initiated by user innovation [55]. Presenting further some famous anecdotal examples, popular sports drink Gatorade was developed by a university football team trainer, the operating system Linux was constructed by computer geeks, and the early versions of protein shampoos and recipes of ready-mixed cakes were developed by house-wives [40].

These user innovators are referred to as lead users. Formally, lead users were defined by von Hippel [61] as users whose “present strong needs will become general in a marketplace months or years in the future” and who “significantly benefit from obtaining the solution to their needs”. The first characteristic refers to users who possess a need prior to other people. This is important because users with a real-world experience of a certain need can articulate the richest information about it [24, 63]. Since they experience it first, they are familiar with conditions that lie in the future for majority of people [63]. Therefore, lead users can serve as a need forecasting laboratory [61]. The second characteristic refers to users with a high benefit expected from solutions to their needs. Because of this high benefit anticipation, lead users are the most likely people to seek for a solution.

Lead users have been involved in NPD of many companies, for example by Airbus, Philips, Nestlé, Gillette, Siemens, LEGO, or 3M [11, 33, 65]. The company probably the most renowned for lead user involvement is 3M. An example from von Hippel et al. [65]’s research about 3M’s lead user involvement is presented to better understand lead user theory. 3M, a high-tech company known for its immense innovation effort, was trying to find out how to prevent infection of patients during surgeries in an economically-friendly way, so hospitals in developing countries would be able to use it. Simultaneously, a new solution was supposed to be more effective than current ones. 3M applied lead user methodology and eventually identified veterinary hospitals as lead users to the infection problem. Veterinary hospitals had presented a strong need about the infection being handled in a cheap and effective way long before regular hospitals (reference to the first lead user characteristic). This is because they take care about "patients" with hairs, no hygiene and mainly no money to pay expensive treatments. Moreover, veterinary hospitals significantly benefit from handling the infection issues (reference to the second lead user characteristic), because almost no animal could get surgery otherwise. 3M adopted the tech-
niques used by veterinary hospitals and introduced a better and cheaper infection prevention for humans.

This section reviews literature about lead users. First, it presents lead users in a theory of innovation diffusion. Second, it describes lead user involvement. Third, it explains a process of lead user identification.

2.1 Lead Users & New Product Adoption

In the innovation diffusion theory, defined by Rogers [52], the diffusion curve shows that every person has different benefits from an innovation adoption (see Figure 2) [51]. The argument comes from the logic that people adopt a product earlier if they expect more benefits from it. In this sense, lead users are individuals with the highest benefit expectation and are therefore ahead of all other groups of adopters [63]. Next to it, the earlier adopter groups are considered more innovative [51].

Noteworthy, lead users are not by definition so called "innovators", the earliest adopter group of the diffusion model. As shown in the figure, they exist before a commercial product is available in the market, therefore before innovators [63]. On the other hand, when a lead user in a certain domain buys a commercial product right after it is introduced in the market, he or she also can be considered an innovator.

Figure 2: Adopters along the Diffusion Curve [65].

Opinion leadership can also be introduced in the context of innovation diffusion. Opinion leaders are individuals with a great influence on others’ behavior and attitude [51]. They play a crucial role in diffusion of innovation, because they accelerate it more than anyone else [32]. Thus, they exist along the whole diffusion curve. Nevertheless, as these individuals are innovative [51], they are expected to be more prevalent in the early stages of the diffusion. Since lead
users are the first to develop novel concepts and products by themselves, they are considered experts. So that, they have a high influence on others [32]. As a consequence, lead users are simultaneously opinion leaders (not vice versa).

Next to that, lead users are rare [63], nevertheless they exist throughout all industries [24]. That includes consumer as well as industrial markets, where they are actually more prevailing [23]. It is interesting to pinpoint that a high lead user prevalence in a certain market may indicate that manufacturers are not interested in that market and users are forced to develop solutions by themselves [17].

2.2 Lead User Involvement

Lead user involvement is a process of transferring knowledge of lead users to firms. It includes activities like the process initialization, lead user identification or product concept development. One of the methodologies that give guidelines how to involve lead users in NPD is Lead User Methodology (LUM) [60]. LUM is depicted in Figure 3. The figure is self-explanatory. Further details are in Urban and von Hippel [60]’s study. This section further describes lead user involvement in terms of its advantages and disadvantages. It is often referred to LUM and not to lead user involvement in general in the following subsections, because LUM is the most widely used methodology to identify lead users [47] and it is well documented.

![Figure 3: Lead User Methodology (LUM) [23].](image)

2.2.1 Advantages

There are various benefits of involving lead users in NPD. It gives a firm for instance a higher chance of developing innovative products [37] in a shorter time and lower development cost [24]. As a consequence, a firm can be better off with higher sales, market share and commercial potential [60], and more satisfied and engaged customers [66]. Last, but not least, lead users are willing to provide information about their inventions [65]. The individual advantages are explained in the next subsections.

It is important to emphasize, that both companies as well as customers can either directly or indirectly benefit from some of the aforementioned advantages. For instance, shorter development time and development cost is a direct advantage for a company as it brings a competitive advantage to the firm. Simultaneously, a customer benefits indirectly by waiting less time to purchase a product with potentially less expenses.
New Product Development

In the first empirical application of lead user involvement, Urban and von Hippel [60] reported that in the printed circuit domain, 87% of lead users are innovative (i.e., building their own solutions) in contrast to 1% of innovative non-lead users. This is underlined with a finding that solutions generated by these innovative users are approximately 30% more novel relatively to the solutions generated by regular users [37]. Lead users develop more innovative solutions, because they are by definition the leading edge of a market and the users who face problems about product deficiencies in more extent than others [65]. Therefore, they are able to provide rich insights about their needs and come up with original solutions. This is in contrast with traditional market research methods (i.e., focus group, need assessment and conjoint analysis), in which regular users are involved. Unlike lead users, regular users are incapable to offer rich insights, because they are constrained with their familiarities [61, 63]. Noteworthy, this is not an issue for incremental innovation, where new products do not drastically differ from the current ones and regular user knowledge does not become out-dated frequently [61, 63]. Moreover, in LUM, a project team gathers information about users’ needs and also potentially innovative solutions to these needs, while only needs are collected in traditional market research methods [65].

Development Time and Cost

Lead users tend to help product teams to come up with product concepts 44% faster than regular users in traditional development methods [24]. This might be because lead user involvement include a lot of parallel activities [24].

Furthermore, development costs are 49% lower [24]. This can be explained in two ways. First, the cost of surveying is much lower, because only around 12 lead users have to be interviewed in comparison to 130 people in traditional methods [24]. Second, ideas provided by lead users are technically less demanding to be developed [24]. To be critical about Herstatt and von Hippel [24]'s results, the authors do not explicitly define what are the traditional development methods, so their findings should be interpreted carefully.

Sales, Market Share and Commercial Potential

Products co-developed by users tend to be more innovative [60]. They appeal more attractive to customers, who are more likely to purchase them. Product concepts developed by lead users have on average eight times higher sales potential than traditionally developed concepts and two times higher market share [37]. Moreover, they also have a high commercial potential. In a study of German surgeons, Lüthje [40] found that one in three surgeons’ inventions are being transferred into marketable products.

Customer Satisfaction and Engagement

Co-developed products tend to increase customer satisfaction and engagement [66]. These effects are argued to arise, because co-created products have a better fit with customer needs [14]. It is important to note that Wayne D. Hoyer et al. [66]’s study concerns general customer in-
volvement domain and does not focus specifically on lead users. However, their finding holds also for lead user involvement as lead users are a sub-group of customers.

**Access to Information**

Lead users are usually willing to provide information about their inventions [65]. Moreover, they are willing to give it free of charge. In case of industrial markets, this is because a lead user and a firm are normally operating in different industries, so they are not competitors [65]. Anti-lock breaking is a good example of this phenomena. It originates from aerospace industry, where airplanes have to stop very quickly. It was later adopted in a non-competing industry, by automobile manufactures [11].

Another argument for the willing provision of information is that a firm pursuing a particular innovation endeavor can help a lead user to develop a needed product [65]. This is especially the case when the innovation is not the firm’s core business (in case of industrial lead users) and the firm rather prefers to buy the innovation from companies specialized in the domain [65].

**2.2.2 Disadvantages**

Lead user involvement also faces some disadvantages. Namely, it is difficult to judge product concept developed by lead users, include lead users in co-development process, and involve them in NPD in a short time.

**Judgment of Product Concepts**

At the time of product concept evaluation, it is hard to judge whether a solution provided by lead users is novel or not. This is because regular users do not experience the same needs as lead users, so they may not see an innovation’s benefits [60]. On the other hand, a different scenario could be that a lead user develops a product concept fitting to his or her needs only, hence it will not become a trend later on [60]. Nevertheless, there are some techniques to remedy these issues [60].

**Involvement in Co-development**

Lead users from consumer markets, particularly in outdoor industry, do not aspire to co-develop products with manufactures in most of the cases [40]. This disadvantage is questionable however, because converse results from basketball domain are found [29]. Note that, lead users might lack of aspiration to actively co-develop products, but are willing to provide needed information [65].

**Identification Time**

Lead users are rare and they represent only a fraction of the population [63]. Therefore, it is difficult to reach them. This conclusion is agreed upon by many researchers. Herstatt and Lüthje [23] reported that identification is the most critical task of LUM. Brem and Bilgram [8] also found that lead user identification is the most time consuming process of LUM. It takes in average 8.5 (SD = 2.7) man-days. Olson and Bakke [46] stated that identifying lead users is
an intricate process. Finally, Schreier and Prugl \[53\] also recognized lead user identification as the major challenge of lead user involvement.

In conclusion, lead user identification is the most important phase of lead user involvement. It is described in detail in the next section.

2.3 Lead User Identification

There are two aspects to consider when searching for lead users \[8\]. First, lead user characteristics need to be identified (i.e., to know who to search for). Second, a lead user search method needs to be selected (i.e., to know how to search). This section further describes both of these aspects in the context of step 2 and 3 of LUM from Figure 3.

2.3.1 Lead User Characteristics

Lead user characteristics can be categorized as general and network specific. General characteristics refer to characteristics describing a user independently of his or her social relationships. Network characteristics describe a user based on his or her social relationships in a network \[20\].

General Characteristics

General characteristics are described in detail in the following paragraphs. A summary is provided in Appendix A.

In a study from sport environment, Schreier and Prugl \[53\] proposed that lead users hold two field related and field independent characteristics. Regarding the field related ones, lead users are knowledgeable users with sound understanding how to do common product related tasks. This allows them to push limits in the target domain, because theoretical expertise is the cornerstone of innovation \[53\]. The second field related characteristic is use experience. Lead users understand ordinary product usage situations, so they can challenge the status quo \[53\].

The field independent variables embraced by lead users are locus of control and innovativeness \[53\]. Lead users are characterized to have a high level of locus of control, thus they are more likely to escape from the prison of their familiarities, deal with new usage situations and challenge flaws in existing products. This lets them master new and difficult situations \[53\]. Schreier and Prugl \[53\] also reported that lead users are more innovative. The innovativeness in this case is defined as the consumption of newness, so lead users are open to new information and they are willing to change the course of their past actions \[53\]. Finally, as a consequence of lead userness, they are among the first ones to adopt new products in a target domain, because they are the first who have experienced a particular need and have the highest benefit from solutions to this need \[53\].

Lüthje \[40\] examined lead user characteristics in the context of outdoor products. He found that lead users face unfulfilled needs and are usually dissatisfied with existing products. These two variables are highly connected. The dissatisfaction with existing products is an indirect measure of unfulfilled needs. This is because the dissatisfaction indicates that a user faces certain unfulfilled needs \[40\]. Similarly to Schreier and Prugl \[53\], Lüthje \[40\] also noted that
speed of adoption is a suitable proxy for lead userness. As depicted in Figure 4, 24.6% of innovative users buy products immediately after they are introduced in the market in comparison to 1.1% of non-innovating buyers. Other typical features of lead users are enjoyment from the innovating process, use experience and product related knowledge [40]. A financial reward is not a predictor of lead userness [40].

Figure 4: Speed of Adoption of Innovating and Non-Innovating Users [40].

Lüthje [40]’s findings are prominent and used by many researchers (e.g., [5, 6]). For instance, Belz and Baumbach [5] built on his work and postulated opinion leadership as a characteristic of lead users.

Jawecki and Füller [29] conducted a study in a basketball online consumer community for shoes and found that the most innovating users are excitement rather than need driven. It is an interesting finding, because it contrasts with von Hippel [61]’s definition of the need driven lead users. According to Jawecki and Füller [29], need driven users express ideas in words, but innovating is not their regular routine. The most innovative users are characterized as individuals who for instance are acknowledged community members, demonstrate impressive drawing skills, and are users with the most innovative content (not with the highest number of posts) [29].

Network Characteristics

Network characteristics describe users’ position in a social network and give specific insights about their social relationships. There are some indicators why network characteristics might help to distinguish lead and non-lead users. For instance, it was reported that lead users are significantly more active in social media than non-lead users [5, 15].

In a recent study, Kratzer et al. [33] motivated by these indicators, inspected social network position of lead users (see Appendix A for an overview of this paper). In particular, they examined whether degree centrality and betweenness centrality might distinguish lead users from the mass. Degree centrality is defined as a number of direct contacts of a user in a network and betweenness centrality as a quantity of a user’s acting as a bridge between different social
crowds. Three studies were conducted in: school, multi-industrial and aerospace setting. Results showed that degree centrality is unrelated to lead userness. Therefore, a person’s number of contacts does not determine one’s lead userness. In contrast, betweenness centrality was found to be associated with lead userness. This is aligned with findings that lead users apply knowledge and experience from different domains to develop solutions to their needs [33]. Kite surfing for instance, combines surfing and hang-gliding domains together [58] or minimal invasive surgery combines modern software applications, new surgical techniques and robotics [34].

2.3.2 Search Methods

The second aspect when identifying lead users is "how to search". Hence, in this section, a review of existing search methods is presented. The search methods are categorized and presented as traditional and social media-based.

Traditional Search Methods

The traditional lead user search methods are screening, pyramiding and signalizing.

**Screening** Screening is the first search technique ever used to identify lead users. It is based on surveying a population sample. Individuals meeting certain criteria are identified as lead users [60]. Screening is normally performed via a questionnaire or a telephone interview [60]. A problem in case of the interview is that it takes too much time. An issue with the questionnaire is that it is based on self-assessment and it is also time-consuming [8]. Moreover, response rate is typically very low [60]. A lot of users need to be approached when screening is employed to find lead users. Only 22 lead users were identified from a sample counting approximately 2,000 people [38].

**Pyramiding** Pyramiding is an approach in which a user with an appropriate expertise is introduced to a problem and asked to refer to a person who has the most relevant knowledge to tackle this problem [65]. The following step is the same like the first one with a difference that the referred individual is approached this time. This is repeated till a lead user is found. The disadvantage of pyramiding is that asking a person to refer to someone more knowledgeable might be a sensitive issue and that it is often undesirable for a company to reveal that it is searching for lead users [8].

**Signalizing** Signalizing is an approach in which a company publicly presents a challenge (i.e., a call for a proposal) and based on users’ responses identifies the most innovative participants [58]. It is a more passive approach, which relies on users’ self-selection [58]. The problem with this method is that it usually requires a permission from a community administrator to conduct the procedure [8]. Additionally, the method’s passivity might be an issue as only a few users respond to calls for proposals [8].
Social Media-based Search Methods

Most of the research about lead users originates from the past century. At that time, the Internet, not even mentioning social media, was not a common tool to use and there was only a limited amount of data available. This has changed. Only between 1995–2015, the internet population increased by a factor of 116 and data grew exponentially with it [28]. This is why it is important to take into account social media when identifying lead users. Moreover, more recent research [5, 16, 33, 58] has shown a potential in analyzing data from social media and indicated that lead users are ordinary habitats in online communities. Finally, as mentioned earlier, lead users are actually significantly more active on social media than regular users [33], which is an important factor for their identification online.

The social media-based search techniques are netnography, crowdsourcing, FLUID, and an approach of Tuarob and Tucker [59].

Netnography Netnography is a qualitative analysis of publicly available data from social media [8]. It can be seen as passive listening of online dialogues without influencing them. This allows to get unbiased information from a user-generated content and is more reliable than using self-assessing methods [8]. The indicators of lead userness in text are for example users’ expressed needs, experiments with products, or attempts to search for a solution. Netnography should not be seen as an approach which solely tries to identify lead users, but also as a way to gain customer insights [8].

Netnography is conducted in five steps [8]:

1. Define search strings for a specific problem domain based on its trends.
2. Identify relevant online communities.
3. Record data (i.e., discussions between users) from a community site.
4. Label data to capture lead user characteristics.
5. Evaluate who is a lead user.

All the steps except of step 4 are self-explanatory. In step 4, need information (e.g., progressive needs), solution information (e.g., ideas or solutions to a problem), and user information (e.g., characteristics like activity and creativity) are captured by manual labeling [8]. The manual labeling of users’ conversations by domain experts is a big downside of netnography, because it consumes a lot of time.

Crowdsourcing Crowdsourcing is based on a public broadcasting of an innovation challenge, which solves a corresponding need gap [8]. There are two kinds of crowdsourcing. The first one requires people to submit an objective solution to a specific problem and the second asks users to submit their subjective opinions or tastes. In both cases, lead users are selected based on the most innovating ideas posted. In contrast to netnography, this method is initiated and actively managed by a researcher [8]. Therefore, it focuses on researcher-to-user dialogues. The big advantage of crowdsourcing is that it is based on a real innovative behavior (i.e., submitting the ideas) [8]. On the other hand, lower amount of users can be reached this way, as they have to proactively sign in the challenge.
**FLUID**  Fast Lead User IDentification (FLUID) is a systematic approach how to identify lead users on social media [47]. It is the proposed first automated approach for lead user identification. In FLUID, an individual’s innovativeness is modeled based on his or her overall behavior on a social media site. This approach takes into account information like frequency how often a user posts messages, relatedness of a user’s messages to a product category of interest, sentiment of messages, time of a user’s messages related to certain objects from a product category of interest (e.g., products), and the degree of a user’s centrality in network [47]. The disadvantage of FLUID is that it approximates a user’s innovativeness based on one’s general behavior and does not take into account the quality of the content well.

**Tuarob and Tucker [59]’s Approach**  The article presented by Tuarob and Tucker [59] avoids the problem which FLUID possesses. It models a user’s innovativeness by inspecting the quality of the content of users’ messages. The quality is measured in terms of prevalence of latent attributes in text. Latent attributes are attributes discussed on social media, but are not present in product specifications. They express users’ unfulfilled needs experienced prior to other people [59], which is one of the two defining characteristics of lead users [61].

The approach of Tuarob and Tucker [59] is visualized in detail in Figure 5. In the figure, either a product or a product category is first defined. The following steps can be done in parallel. Describing the left branch, product specs from manufactures are manually found. From these documents, product specification attributes are extracted using Revminer, a review extraction system proposed by Huang et al. [25]. The process of the extraction consists of a manual initialization by several specification attributes and their values. Based on this input, Revminer finds templates in data and extracts new specification attributes with their corresponding values.

The right branch starts with a manual definition of search terms for a product or a product category on social media. For instance, a search term for a product Samsung Galaxy S6 Edge Plus can be "Galaxy S6" or "Galaxy S6 Edge". Then, product attributes and their values are extracted from social media in the same way like specification attributes. These attributes are denoted as user-discussed attributes.

In the following step, product latent attributes are identified. They are computed by subtracting the sets of user-discussed and specification attributes. The weight of latent attributes is determined by an algorithm similar to Term Frequency (TF) and Inverse Document Frequency (IDF). The rationale of using TF is that when a particular attribute is missing in current product specifications, many conversations about this attribute are expected to be present on social media [59]. TF captures this phenomena. From the other side, when a latent attribute is associated with a few products only, it is assumed that an attribute is uncommon and innovative. That is quantified by IDF. The weighting enables to rank attributes and also further rank users by their lead userness. The more a user talks about important latent attributes, the more he or she is considered to be a lead user.

The proposed method is tested by conducting a case study in the domain of mobile phones. Waterproofness is an example of one of the extracted latent attributes. As shown in Figure 6, this feature had been discussed long before the first waterproof phone (i.e., Sony Experia Z) was introduced. Although the results of Tuarob and Tucker [59] are rather anecdotal and not verified by domain experts, they indicate that using product attributes can help to identify lead users and their insights.
3 Metrics

Metrics represent scores based on which users can be ranked. They are hypothesized to capture some of the characteristics of lead users. Hence, lead users are expected to score higher on the analyzed metrics, because they possess these characteristics in greater degree than ordinary users. To illustrate this, a chance of selecting a lead user from a group of people who have domain product knowledge is higher than from a group of randomly selected users. This is because a randomly selected user might not be knowledgeable about products and consequently not be a lead user, since lead users possess product knowledge by definition. Inspecting various metrics in relation with lead userness is beneficial, because different characteristics of lead users can be captured and in turn different subgroups of lead users can be potentially identified.

Selected metrics are theoretically explained in this section. Metrics’ definitions and reasoning why they are suggested to be related with lead userness are presented for every metric. Metrics are categorized into network, text and attribute metrics.

3.1 Network Metrics

Network metrics are metrics built upon a discussion network of community users. A node in a network represents a user. Nodes are connected by links. A link from user A to user B is established when user A commented on a message of user B (see Figure 5). A link from A to B conveys different information than a link from B to A, so links are directed.
Different network metrics express a user’s position in network from different perspectives. In this section, eight various network metrics are introduced. They are indegree centrality, outdegree centrality, domain-interest, betweenness centrality, closeness centrality, pagerank, hub, and authority.

3.1.1 Outdegree Centrality

Outdegree centrality can be expressed as a number of outcoming links from a node. It refers to an amount of messages posted by a user in a particular thread.

Lead users are hypothesized to score higher on outdegree centrality, because they are knowledgeable [53] and involved people [5], who are likely to understand and react on messages of others, thus having a higher outdegree centrality.

Outdegree centrality can be formally defined as:

\[
C_o(i) = \deg_o(i)
\]  

(1)

where \(\deg_o(i)\) is an amount of outcoming links from node \(i\).

3.1.2 Indegree Centrality

Indegree centrality can be expressed as a number of incoming links to a node. It refers to an amount of direct responses to a user’s messages in a particular thread.
Lead users are hypothesized to score higher on indegree centrality, because they are innovative people with outstanding ideas [53] and also opinion leaders [5], who highly influence the mass. Therefore, their posts are expected to be highly discussed by others, thus lead users are presumed to have a higher indegree centrality.

Indegree centrality can be formally defined as:

\[ C_i(i) = \deg_i(i) \]  

where \( \deg_i(i) \) is an amount of incoming links connected to node \( i \).

### 3.1.3 Domain-Interest

Domain-interest is a newly defined metric. It refers to a proportion between messages of a domain of interest and all a user’s messages in a whole community forum. Judging a user’s activity based on his or her overall behavior on a whole social media site can be more informative than just looking at one topical thread like all other metrics do.

Lead users are hypothesized to score higher on domain-interest. As mentioned, lead users are involved people in a particular product domain [5]. It is foreseen that the most of their attention in a community forum is focused on the domain of interest. Thus, lead users are expected to have a higher domain-interest.

Formally, domain-interest is defined as:

\[ C_{di}(i) = \frac{\sum_{m \in B} 1}{\sum_{m} 1 + C} \]  

where \( M_i \) is a set of the latest \( Z \) messages of a user \( i \) from whole social media site no older than \( T \) days, \( B \) is a subset of \( M_i \) referring to messages related to a domain of interest. Note, that the denominator is regularized by a constant \( C \). It prevents users with only a few messages, which are mainly about a product category of interest, being ranked high.

### 3.1.4 Betweenness Centrality

Betweenness centrality of a node can expressed as an amount of nodes going through the given node in order to reach other nodes in the minimum number of hops.

Lead users are hypothesized to score higher on betweenness centrality. This is because they are likely to create innovations with respect to their needs [53], which is partly expressed by betweenness centrality. The reasoning goes like this. Innovation often occurs by a combination of knowledge from different domains [34]. Since users scoring high on betweenness centrality have an access to different knowledge domains, because they connect different social crowds, it is easier for them to innovative and supposedly create innovations just like lead users. Thus, lead users are expected to have a higher betweenness centrality. In fact, a positive relation between this metric and lead userness has been reported by Kratzer et al. [33].

Formally, betweenness centrality is defined as:

\[ C_b(i) = \frac{\sum_{j<k} g_{jk}(i)}{g_{jk}} \]  

where \( g_{jk}(i) \) is the number of shortest paths between nodes \( j \) and \( k \) that go through node \( i \).
where $g_{jk}$ is an amount of the shortest paths connecting nodes $j$ and $k$, $g_{jk}(i)$ is an amount of the shortest paths between nodes $j$ and $k$ in which $i$ appears, and $N$ is a total amount of nodes [1].

### 3.1.5 Closeness Centrality

Closeness centrality expresses how close a user is to the center of a network [1]. It refers to the length of an average shortest path between a given node and all other nodes in a network.

Lead users are hypothesized to score higher on closeness centrality. The reasoning is similar to the arguments of betweenness centrality. A user central in a network might have an easier access to groups with various backgrounds, combine knowledge, be innovative, and have a higher chance of being a lead user.

Closeness centrality is defined as:

$$C_c(i) = \left[ \sum_{j=1}^{N} d_{ij} \right]^{-1}$$

where $d_{ij}$ is the minimum distance between nodes $i$ and $j$ and $N$ is a total amount of nodes [1].

### 3.1.6 Pagerank

Pagerank originates from hyperlink analysis and is commonly used to assess the graph structure of the Web [41]. It interprets a link from node $i$ to node $j$ as a vote of $i$ to $j$ [49]. The more votes $j$ receives from other important nodes, the more important $j$ is.

Lead users are hypothesized to have a higher pagerank, because pagerank is related to a user’s expertness [30, 48], which is one of the characteristics of lead users. Specifically, it is an expertness in terms of product knowledge and product use experience [53].

Pagerank is formally defined as:

$$PR(i) = \frac{\lambda}{N} + (1 - \lambda) \sum_{j \in B_i} \frac{PR(j)}{L_j}$$

where $\lambda$ is a probability of teleporting to another page, $N$ is a total number of pages, $B_i$ is a set of all nodes linking to $i$, and $L_j$ is a number of outgoing links from $j$ [49].

### 3.1.7 Hub and Authority

Hub and authority metrics are two different metrics. They are presented together, because their definition is circular. A good hub node is designated to point to many authoritative nodes in a network [49]. Likewise, a good authority node is designated to be pointed by many good hubs [49].

Just like pagerank, hub and authority metrics originate from the field of hyperlink analysis and can be also used to identify influential users (e.g., [48]). Hence, lead users are expected to have a higher hub and authority scores.

Hub metric is formally defined as:

$$h(i) \leftarrow \sum_{j \in B_i} a(j)$$
where $B_i$ is a set of pages to which $i$ links to and $a(j)$ is authority score. Authority score is defined as:

$$a(i) \leftarrow \sum_{j \in D_i} h(j)$$  \hspace{1cm} (8)

where $D_i$ is a set of pages which links to $i$. These equations are run iteratively with arbitrary initial values till they converge [49].

### 3.2 Text Metrics

Text metrics intend to capture lead usereness via analyzing content of users’ messages. The presented text metrics are latent-attributes and specification-attributes-in-time. They are both based on extraction of product attributes from unstructured text. So that, a review about relation extraction, a super-category of attribute extraction, is presented in Appendix B.

#### 3.2.1 Latent-attributes

Latent-attributes metric aims to capture one of the defining characteristics of lead users, the degree of user’s unfulfilled needs about missing features of a product (i.e., latent attributes) experienced prior to other people [59]. The more a user talks about latent attributes, the more he or she is considered as a lead user. Latent attributes are attributes which are discussed on social media, but are not present in product specifications [59]. A latent attribute can be for instance solar charging of a smartphone, because it is discussed on social media with an association with smartphones and because there is currently no phone with this attribute in the market. Next to latent attributes, a term specification attributes is also further used in the text. Specification attributes are attributes of existing products. LCD panel or bluetooth are for instance specification attributes of Samsung Galaxy S7.

In LUM, lead users are defined with respect to trends [61]. An example of a trend in the area of printed circuits boards is density of chips placed on a board [60] or it can be a noncorroding and lighter pipe hangers in the area of equipment and material used in construction [24]. A trend also expresses the degree of user’s unfulfilled needs, however it does not refer to the same concept as latent attributes. A latent attribute can be an attribute in need for any size of market while a trend refers to attributes in need for a majority of market. Therefore, it is proposed that trends are a subgroup of latent attributes\(^3\).

Lead users are hypothesized to have a higher latent-attributes score. This is because unlike ordinary users, lead users face many unfulfilled needs about products from the domain [61]. The more unfulfilled needs lead users possess, the more they are expected to talk about latent attributes, thus scoring higher on latent-attributes metric.

Latent-attributes metric is defined as:

$$La(i) = \sum_{m} \sum_{a \in m} 1$$  \hspace{1cm} (9)

\(^3\)Considering only trends that can be translated into physical objects (i.e., attributes).
where $M_i$ is a set of all messages of a user $i$ and $A_l$ is a set of all latent attributes a message $m$ contains. The metric simply represents a count of all latent attributes in a user’s messages.

Parameters of this metric are latent attributes. To determine them, in this section, two techniques for latent attribute extraction are introduced. The first is a simplified version of template matching used by Tuarob and Tucker [59], which establishes a baseline for a latent attribute extraction. The second is a new technique. It proposes to extract latent attributes using a word embeddings of Word2vec framework.

**Extraction via Template Matching**

A template matching technique is proposed to establish a baseline for latent attribute extraction. It is based on the work of Tuarob and Tucker [59]. Further in this section, there are explained differences between the proposed and Tuarob and Tucker [59]'s approach along the methodological framework from Figure 5.

Step 1 and step 2 of the methodology, retrieving product specifications and extracting product attributes, are intended to be automated by using a service called Semantics3. This service provides product specifications across many product categories. It is evaluated whether Semantics3 is a suitable source for attribute retrieval (see Appendix C for details). It turned out that this service can in general retrieve specification attributes with high precision, but with a low recall. Therefore, Semantics3 is not used to retrieve product specifications and specification attributes and these steps are proposed to be done manually.

Step 3, retrieving product names, is suggested to be done manually like Tuarob and Tucker [59] did.

Step 4, extracting product attributes from social media, is proposed to be performed by Brin [9]'s template matching system DIPRE. Extraction system Revminer, which was employed by Tuarob and Tucker [59], is not used because of a more complex implementation, which would not fit the limited time scope of the project. To describe DIPRE, in the context of attribute extraction, templates or in other words patterns are strings between product names and product attributes. DIPRE extracts new attributes using seed examples of both of the entities. When two product names or attributes match a pattern, the longer expression in terms of word count is used. For instance, when "iphone 6s plus" and "iphone" both match a pattern, the first one is used. Concerning the setting of the template matching, left and right contexts of product names and attributes are not captured to simplify the implementation. Following on the literature review about relation extraction, DIPRE needs to be set with a several parameters. They are a minimum number of occurrences of a pattern $T_n$, a number of iterations $T_i$ in which the template matching runs. Furthermore, there is a minimum pattern specificity $T_p$ and a minimum number of occurrences $T_o$ for a term to be identified as a candidate attribute.

It is not expected that an approach using template matching with specification attributes as an input will yield many latent attributes. This is because template matching approaches are based on data redundancy and patterns need to be repeated to be successfully extracted. As latent attributes are assumed to be mainly discussed by lead users, who are rare, the latent attributes are foreseen to be presented in much lower frequencies than specification attributes. Thus, they might be considered noise.
Extraction via Word2vec

The second proposed approach for latent attribute extraction is based on word embeddings. In word embeddings, a term is defined by its context. The context is represented by neighboring terms of the given term. The size of the neighborhood is determined by a parameter \( w \), called window length. For instance, in a sentence like “A queen regnant possesses and exercises sovereign powers.”, a word "regnant" is represented by neighboring words "queen" and "possesses" when \( w = 1 \). When two words have a similar context, they are considered to be semantically similar. According to distributional hypothesis, semantically related terms appear nearby each other in a created vector space \([57]\). Such vector space embeds words into vectors. It can be created for instance by Word2vec framework.

The embeddings of Word2vec encode many linguistics patterns, which can be represented as linear translations \([42]\). For instance, \( \text{vec("king")} - \text{vec("man")} + \text{vec("woman")} \) is close to \( \text{vec("queen")} \), or \( \text{vec("Russia")} + \text{vec("river")} \) is close to \( \text{vec("Volga River")} \) \([42]\). Word2vec can be also used to search for similar objects. That can be formally expressed as \([35]\):

\[
sim(i, N) = \arg \max_{j^* \in V}(\cos(i, j^*))
\]

where \( V \) is a set of all terms in a vector space, \( N \) is a number of the most similar terms to be returned, and \( i \) is an input term(s). When multiple terms are used as an input, their representation is averaged and they are treated as one word. Note that this equation is further referred to as the Word2vec similarity query.

Word2vec outperforms parsing approaches, such as template matching, in extraction of semantic relationships \([4]\). Therefore, it is expected that using Word2vec will yield more precise results than the template matching technique. Moreover, two of the three introduced Word2vec approaches tackles the issue of using specification attributes as input for latent attribute identification. The introduced approaches are No Latent Context, Implicit Latent Context and Explicit Latent Context, described in the following sections.

No Latent Context  This approach translates Tuarob and Tucker \([59]\)’s methodology into setting of word embeddings. It can be expressed as \( \text{sim}(i, N) \), where \( i \) is a set of specification product attributes and \( N \) is a number of the most similar terms to be returned (i.e., a cut-off value). It is expected that the query will result mainly in a set of specification attributes, because similarly like in the template matching, it is expected that most of the specification attributes will not be discussed in the same context like latent attributes (i.e., latent context). This approach is examined to see if this presumption is correct.

Implicit Latent Context  A possible remedy for the former approach might be using a subgroup of specification attributes as an input for the similarity. Let split specification attributes into latent specification attributes and feedback specification attributes depending on the time when they are talked about on social media. Latent specification attributes are discussed prior to their release dates and feedback specification attributes are attributes discussed after their release dates. Latent specification attributes refer to at the time non-existing attributes, which later become implemented into products. Because of that, they are assumed to satisfy majority of the market. Therefore, unlike latent attributes, they refer to the same concept like trends
in LUM. So that, latent specification attributes are deemed to be more useful for lead user identification than latent attributes.

In this approach, latent attributes extraction is established by the similarity query \( \text{sim}(i, N) \), where \( i \) is a set of latent specification attributes. The query is expected to result in a set of other latent specification attributes and a set of latent attributes. This is because both of the sets are semantically similar.

An example of segmentation of waterproofness attribute into latent and feedback specification attribute is shown in Figure 8. The figure shows that the segmentation also consists of a time span \( s \), which is subtracted from an attribute’s release date. The new point in time determines the attribute group. Attributes discussed prior to this point are latent specification attributes and after are feedback specification attributes. The time span is introduced, because there might be an announcement or a leak of information about a product prior to an attribute release. Therefore, users can start to talk about it while not expressing unfulfilled needs. On the other hand, even if a user talks about an attribute right after an announcement, he or she exhibits a behavior of an innovator (the adoption group), the second most innovative behavior given the diffusion model of innovation \[51\]. That means that it is not perceived as a problem if the span is too small and does not cover an actual announcement date.

![Figure 8: Segmentation of Waterproofness into Latent and Feedback Specification Attributes.](image)

Word2vec needs to be provided with a sufficient amount of examples to correctly capture semantic relationships. This might be a problem, because similarly like latent attributes, frequency of latent specification attributes is expected to be low.

**Explicit Latent Context** To address the potential issue of too few occurrences of latent specification attributes, the context of latent attributes is proposed to be modeled explicitly. Because linear transformations of vectors are meaningful in word embeddings \[42\], this might be done by a simple addition of a vector representing a latent context and a vector representing specification attributes.

The latent context is incorporated into the similarity ranking in a way that every term returned from the similarity query is weighted with a similarity of a context word(s). This can be formally expressed as:

\[
\text{sim}_{\text{context}}(\vec{i}, N, C) = \text{sim}(\vec{i}, N)^T \times \vec{C} \tag{11}
\]

\[\text{Actually, in Rogers } [51]'s model, lead users are not included, so innovators are considered as the most innovative group.\]
where $i$ and $N$ are the parameters of the similarity query of Equation 10 and $C$ is a matrix of vectors of context words.

### 3.2.2 Specification-attributes-in-time

Specification-attributes-in-time is a newly defined metric. It follows on the idea of segmenting specification attributes into latent and feedback categories. The segmentation is done in a continuous way, which introduces more variance to the metric’s scale. Specification-attributes-in-time scores users based on prevalence of weighted specification attributes mentioned in their messages. The weight is assigned to each attribute depending on an attribute’s release date relatively to the time when a message containing an attribute is posted. The attribute weighting is done according to a following function (see Figure 9):

$$w(x) = \begin{cases} 
10 & x \in (-\infty, 180) \\
10^{\frac{x}{360}} \times (10^{\frac{1}{360}})^{-x} & x \in [-180, 360] \\
1 & x \in (360, \infty) 
\end{cases}$$

(12)

where $x$ is the time of an attribute being discussed relatively to the attribute’s release date. The formula is meant to be similar with logistic function used to model innovation diffusion. It was derived by simple mathematical operations given several assumptions. It is presumed that the importance (i.e., weight) of an attribute when it is a latent specification attribute is 10 fold higher than when it is a feedback specification attribute. This determined the $y$-axis range. Secondly, it is presumed that the importance of an attribute varies in the range of 180 days prior to the attribute’s release and 360 days after the release and that the relationship is exponential, like the diffusion curve. Finally, the exponent of the function was determined to be -1, to represent a gradually increasing slope.

![Figure 9: Attribute Weighting Function $w(x)$ of Specification-attributes-in-time.](image)

Specification-attributes-in-time metric assigns a score to every user as a sum of weighted attributes, which a user has posted. It is formally expressed as:

$$Sa(i) = \sum_{m}^{M_i} \sum_{a \in m}^{A} w(m_t - a_t) \ln(d_i)$$

(13)

where $M_i$ is a set of messages a user $i$ has posted, $A$ is a set of attributes a message $m$ contains, $m_t$ is the time when a message was posted and $a_t$ is the time when an attribute was released,
and \( d_i \) is a total amount of messages of a user \( i \). The denominator intends to normalize the score, so that the metric will not be simply biased towards users sharing many messages.

Lead users are hypothesized to score higher on specification-attributes-in-time metric, because of several reasons. First, this metric is inspired by diffusion of innovation model, in which the earlier adopters are considered more innovative [51]. Reflecting the diffusion model, the attributes posted earlier relatively to their release dates are consider more innovative and have a higher weight. Since lead users exist even prior to the first adopters [65], they are the most innovative group, so they are expected to score higher than ordinary users. This is aligned with a finding that speed of adoption is a suitable proxy to assess lead userness [53]. Second, because lead users have unfulfilled needs [61] and are knowledge [53]. The attribute weighting captures these characteristics by taking into account latent and feedback specification attributes respectively.

There are several advantages of specification-attributes-in-time. The usage of specification attributes avoids relying only on sparse data like latent-attributes metric does. Next to it, using this metric might enable to skip trend identification in LUM, which can take up two months and cost up to $9,000 [24]. This is because the attribute weighting implicitly takes into account latent specification attributes, which are considered trends. Finally, it is proposed that since the weighting function scores users similarly like the model of innovation diffusion, it quantifies a behavior of different adopter groups and might also enable their identification.

Specification-attributes-in-time also possess a disadvantage. It is not certain that a user which has talked in past about latent specification attributes is still active in the product domain and that his or her status of a potential lead user has not "expired".

Specification-attributes-time is a new metric in the context of identification of lead users. The newly introduced part is the weighting function. A baseline is created to inspect whether the attribute weighting is useful for lead user identification. The baseline omits using the attribute weighting and is defined as a plain count of specification attributes.

### 3.3 Attribute Metrics

Attribute metrics are metrics derived from meta-data of messages. Attribute metrics include votes metric.

Votes is metric that refers to an aggregated count of a user’s votes. It represents a likeability of a content of a given user. Lead users are hypothesized to score higher on this metric, because the content of their messages is expected to be popular, as they are able to provide expert opinion in a domain [53] and possibly help other people [29]. Moreover, lead users are people with an access to broad community resources [16], so community users might tend to like their messages more, because there is an existing relationship between them and lead users. On the hand, ideas of lead users might not be appreciated by community users, because general market place lags in recognizing innovative ideas [61].

Formally, this metric is defined as:

\[
v(i) = \sum_m m_{votes}
\]  

(14)
where $M_i$ is a collection of messages of a user $i$ and $m_{votes}$ is an amount of votes a message $m$ received.

4 Case Study

The goal of the case study is to inspect which metrics can help to find lead users in a product-oriented forum. The case study aims to show a path towards a fast lead user identification.

This section first introduces selected product domain and data corpus. Then, computation of metrics is described. Subsequently, it is explained how ground-truth labels of lead userness of community users were obtained. Afterwards, the relationship of the metrics and lead userness is assessed and results of the case study are presented. Finally, a tool LUSA to help domain experts to identify lead users is explained.

4.1 Domain & Data

The domain chosen for this research is the smartphone domain. It is a high tech and a high involvement product domain. It was selected because lead users are the most prevailing in such domains [5], so there would be a higher chance that rare lead users are involved in the analysis.

Selected data source is the social media site Reddit. Reddit was chosen due to its vivid community activity, strong accumulation of knowledgeable users and mainly due to the possibility of accessing all data free of charge. In particular, data from a thread about iPhones [5] were retrieved. Note, that a thread in Reddit has a similar structure like a regular community forum.

Data were collected from 2008-01-26, the time of the first post in the iPhone thread, and 2016-07-20, the time when the research was being done. There were 1,003,769 messages [6] retrieved from 113,030 users. The corpus consists of 32,289,631 words, that is in average 32.17 words per message. As shown in Figure 10, the distribution of the amount of the users’ messages is skewed to the left. That means that most of the users have posted only a few messages.

The users with less than 15 messages were dropped from the analysis due to low activity. It would be difficult to assess them based on such limited amount of information. This resulted in a set of 10,731 users.

Finally, it was recorded in which other threads on Reddit are the users the most active.

4.1.1 Data Preprocessing

The messages were lowercased. All URLs and special characters (apart from underscore and hashtag) were removed. Moreover, consecutive white spaces were substituted by a single space. Multi-word expressions (i.e., bigrams and trigrams) were formed by merging frequently co-occurring terms together. That was done via Gensim library which implements the approach of Mikolov and Dean [42].

---

5http://www.reddit.com/r/iphone

6A message in the forum is considered to be a heading of a thread or a comment in a thread.
4.2 Computation of Metrics

This section introduces a setting in which the selected metrics were computed.

4.2.1 Network Metrics

Network metrics were computed from messages of all 113,030 users of the iPhone thread. The characteristics of the social network are presented in Table 1. The amount of nodes refers to the (unfiltered) amount of users. There is 702,236 links between the users; in average, 6.21 links per user. Network diameter, the shortest distance between the most distant nodes in a network, is 16. The average shortest path length between any two nodes is 14.87 links. The network is not connected very densely.

Table 1: Characteristics of the Analyzed Network.

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Edges</th>
<th>Avg. Degree</th>
<th>Network Diameter</th>
<th>Avg. Path Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>113,030</td>
<td>702,236</td>
<td>6.21</td>
<td>16</td>
<td>14.87</td>
</tr>
</tbody>
</table>

The parameters of the network metrics are shown in Table 2. Most of the network metrics were calculated by a software tool Gephi [18] and did not require any input to be defined. Only two metrics need to have parameters set. They are pagerank and domain-interest.

Parameter $\lambda$ of pagerank expresses a probability of teleporting to another page. It is specified to be $\lambda = 0.15$, which is in a range of usual values [49].

Parameter $B$ of domain-interest is a set of threads about a product category of interest. To compose this set, threads in which are the top 200 ranked users on each metric active were examined by the author. Threads in which the users posted in average less than two messages were skipped. The selected threads about smartphones and related gadgets are displayed in Table 3. The regularizing constant $C$ was experimentally chosen to be $C = 20$. An amount of messages taken into analysis $Z$ and the earliest time $T$ from when these messages are posted were arbitrary chosen as $Z = 500$ messages and $T = 180$ days.
<table>
<thead>
<tr>
<th>Metric</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outdegree centrality</td>
<td>-</td>
</tr>
<tr>
<td>Indegree centrality</td>
<td>-</td>
</tr>
<tr>
<td>Domain-interest</td>
<td>Set $B$ is defined in Table 3. $C = 20$, $Z = 500$, $T = 180$</td>
</tr>
<tr>
<td>Betweenness centrality</td>
<td>-</td>
</tr>
<tr>
<td>Closeness centrality</td>
<td>-</td>
</tr>
<tr>
<td>Pagerank</td>
<td>$\lambda = 0.85$</td>
</tr>
<tr>
<td>Hub</td>
<td>-</td>
</tr>
<tr>
<td>Authority</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3: Selected Threads about Smartphones on Reddit.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>iphone, apple, tmobile,</td>
<td>Android, redditmobile, mobilerepair, windowsphone, GalaxyS7,</td>
</tr>
<tr>
<td></td>
<td>applehelp, iOSBeta, ios,</td>
</tr>
</tbody>
</table>

4.2.2 Text Metrics

In this section, parameters of text metrics are presented.

Latent-Attributes

The parameters of this metric are latent attributes. Latent attributes were identified via the template matching and Word2vec extraction techniques.

Extraction via Template Matching  Several parameters need to be set to perform latent attribute extraction by the simplified template matching of [59]. The set of specification attributes was manually extracted from EveryMac [13] (see Appendix D). The set of product names was manually assembled from popular smartphone names from GSMarena [21] (see Appendix D). The parameters for DIPRE’s template matching are in Table 4. There were selected experimentally. To clarify surprising setting of $T_i = 1$, the new extracted attributes have a very low precision even after only one interaction. Running the algorithm again would introduce a lot more noise.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_n$ - min. number of pattern occurrences</td>
<td>3</td>
</tr>
<tr>
<td>$T_p$ - min. pattern specificity</td>
<td>2</td>
</tr>
<tr>
<td>$T_o$ - min. number of occurrences to identify a term as a candidate attribute</td>
<td>40</td>
</tr>
<tr>
<td>$T_i$ - number of iterations</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4: Parameters of DIPRE.

The template matching was performed given the presented setting. Identified terms were evaluated in terms of precision@100 judged by the author. First, it was evaluated whether the identified terms are specification attributes of smartphones. A term was labeled as a specification attribute when it is either a name of an attribute or its values (e.g., zeiss, a value of
attribute camera). Second, it was assessed whether the extracted terms can be considered latent attributes. This is not a straightforward task, because no context information specifying the terms is provided. For instance, it is not certain whether a term projector is a latent attribute of an iPhone or it ranks high because of another reason. A term is labeled as an latent attribute if it can meaningfully expand an iPhone.

From the first 100 terms ranked on the occurrence frequency $T_o$, only nine terms are specification attributes and zero are latent attributes. The simplified template matching appears not to be a suitable approach for extraction of latent attributes.

**Extraction via Word2vec** The parameters of latent attribute extraction via Word2vec are stated as follows. The Word2vec model is trained with a window length of $w = 5$. A smaller window length is selected in order to capture more of functional (rather than topical) relations [36]. The architecture used is CBOW due to the smaller size of the corpus. The same set of specification attributes like for the template matching is used for No Latent Context, Implicit Latent Context and Explicit Latent Context approaches. Additional parameters are mentioned in Table 5.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Latent Context</td>
<td>-</td>
</tr>
<tr>
<td>Implicit Latent Context</td>
<td>$s = 30$</td>
</tr>
<tr>
<td>Explicit Latent Context</td>
<td>$N = 5000$, $C = {[\text{wish, would, desire}], [\text{samsung, nokia, android}], [\text{wish, would, desire, samsung, nokia, android}]}$</td>
</tr>
</tbody>
</table>

In Implicit Latent Context approach, the segmentation of specification attributes is done by pre-processing text corpus and attaching two different post-fixes to each attribute depending on the time when they were discussed. The length of span $s$ is $s = 30$ days. The bigger the $s$, the lower the chance that a user posts a message with a latent attribute, because of prior announcement or an information leak. On the other hand, this also makes a set of latent specification attributes smaller. Therefore, the length of $s$ is set to be relatively short, to 30 days.

In Explicit Latent Context approach, a number of attributes $N$ to be returned from the similarity query was experimentally set to $N = 5000$. Vectors of context words $C$, expressing a latent context, were selected arbitrary based on a context in which latent specification attributes are mentioned in the corpus. This context includes sentences expressing desires like “I wish my phone would have a solar panel.”, or comparing an iPhone with a competition while pointing at missing attributes like “Samsung has a micro sd card, I do not understand why iPhone does not.”, or referring about question about attributes in need like “Do you know when iPhone would have a replaceable battery?”. The most straightforward latent contexts to be expressed by explicit words are perceived to be the desire and the competition contexts. The word "desire" is described as “to wish or long for; crave; want” and “to express a wish to obtain; ask for; request”[10]. So that, the words representing the desire context are selected to be "wish", "would" and "desire". The competition context can be expressed by the biggest competitors of iPhone. Therefore, the words selected for the competitive context are "Samsung", "Nokia"
and "Android". The groups of context words were used individually (as a group) as well as in combination with each other to query latent attributes.

The extraction via Word2vec was performed given the presented setting. The evaluation of candidate attributes, as shown Table 6, was done in the same way like the evaluation of the template matching. Table 6 shows that No Latent Context approach is feasible for identification of specification, but not latent attributes. From the highest 100 ranked candidate attributes on the semantic similarity provided by similarity query, 76 are specification attributes and only 2 are considered latent attributes. Implicit Latent Context approach shows an improvement in the extraction of latent attributes. There are 6 latent attributes in the result set. Explicit Latent Context approach enables to extract 13, 7 and 9 latent attributes using the desire context, the competition context and their combination respectively. Noteworthy, the extracted latent attributes by the implicit latent and the explicit latent contexts do not overlap.

<table>
<thead>
<tr>
<th>Latent Context</th>
<th>Category of Attributes Used</th>
<th>Spec Attrs. - Precision@100</th>
<th>Latent Attrs. - Precision@100</th>
<th>Identified Latent Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>Specification</td>
<td>0.76</td>
<td>0.02</td>
<td>amoled, touchpad</td>
</tr>
<tr>
<td>Implicit</td>
<td>Latent specification</td>
<td>0.52</td>
<td>0.06</td>
<td>tri core, 14nm process, polarizing filter, gaming controller, flashable, iris scanner expandable storage, microsd slot, stylus, removable battery, water resistance, physical keyboard, wireless charging, dslrs, physical button, optical zoom, sd card, amoled screen, ps3 controller</td>
</tr>
<tr>
<td>Explicit</td>
<td>Specification</td>
<td>0.39</td>
<td>0.13</td>
<td></td>
</tr>
</tbody>
</table>

The results are presented with the desire context words, which achieved the best precision.

There were 20 latent attributes extracted together from all of the approaches. This can be interpreted that 20 ideas how to expand an iPhone can be provided. This is a high amount, especially comparing it with zero ideas extracted by the template matching approach. Implicit Latent Context approach might extracted only a few latent attributes, because the count frequency of latent specification attributes is too low. There is 848 latent specification attributes in the corpus. That is 171 times less than the frequency of feedback specification attributes which is 145,345. This might explain why No Latent Context approach extracted mainly specification attributes. The feedback specification attributes strongly dominated the similarity query, so mainly feedback specification attributes were provided. The frequencies of individual latent and feedback specification attributes are displayed in Appendix E.
**Selected Latent Attributes**  The attributes used as an input for latent-attributes metric are all the latent attributes identified with the proposed approaches and a few additional ones, which were discovered while exploring the data. The extracted attributes like projector were not used, because they are ambiguous and can be easily used in another context. The latent attributes used in the experiments are displayed in Appendix F. The frequency count of all selected latent attributes is 1875, which is 0.17 latent attributes per user. This might be too low to differentiate users in terms of their unfulfilled needs and eventually in terms of their lead userness.

**Specification-attributes-in-time**

The only parameter to specify this metric and its baseline is a set of specification attributes. The same set of attributes like in the template matching is used.

**4.2.3 Attribute Metrics**

No parameters for attribute metrics need to be determined.

**4.3 Obtaining Labels**

A traditional lead user identification technique, screening, was used to obtain information about users’ lead userness. Screening was performed based on surveying a population sample in a form of a questionnaire. The questionnaire, as shown in Appendix G, measures lead userness with respect to a product category (of smartphones) like e.g. Belz and Baumbach [5] did and not with respect to trends like it is done in LUM [61]. The questionnaire consists of six constructs referring to the characteristics of lead users: ahead of trend (AOT), high expected benefits (HEB), product related knowledge (PRK), involvement (INV), opinion leadership (OL), and community-based resources (CBR). AOT and HEB are the defining characteristics of lead users [61]. The others are correlated characteristics. Each construct consists of three items.

The questionnaire is built from several different sources (i.e., [5, 16, 19]). This is because the questionnaire of Goldsmith and Witt [19] consists of only a few constructs and might not fully capture lead userness. The other questionnaires [5, 16] contain items like “I regularly prepare dishes which contains sustainable food.”, which would be difficult to convert into the smartphone domain terminology. The items of the used questionnaires were adjusted to fit the smartphone domain. For some items, only a change of a domain name was required. For others, more effort had to be done. For instance, an item “Sustainable food matters to me.”, which measures involvement, was changed into “The advancement of smartphones matters to me.”, because it would not make sense otherwise. A commonly measured lead user characteristic use experience is not used, because it is presumed that majority of people use phone daily and everybody would score high. Thus, this would not allow to differentiate between lead and non-lead users.

Next to the lead userness measurement, the participants are asked to indicate whether they have a smartphone and consequently which brand.

The order of the items is randomized for every user to prevent ordering bias.
4.3.1 Pilot Study

A pilot study was performed to assess the questionnaire’s reliability. There were 41 people, who participated in the study. The participants are mainly TU/e students.

Cronbach’s alpha of all the constructs except of PRK is higher than the acceptable level of 0.7 (see Appendix for details). That reflects the comments which the participants made about the questionnaire. Mostly, all the questions were clear, however two items of PRK were indicated to be vague. The degree of the repair of item “I can repair my own mobile phone.” was not clear (e.g., is it SW or HW?). The second item “I have difficulties making technical changes to my mobile phone.” was dubious in terms of the degree of a technical change (e.g., is it referring to changing a background or replacing a CPU?). Therefore, these items were further specified in the main study as “I can usually fix both software and hardware issues of my phone.” and “I have difficulties making software and hardware changes to my smartphone.” respectively. Apart from the third item of PRK (i.e., PRK3), item-total correlations of each construct are higher than 0.7. That suggests in general a consistent behavior of each construct. Omitting PRK due to its low reliability, confirmatory factor analysis suggests a relatively good fit for data, $p = 0.150$, $CFI = 0.941$, $TLI = 0.925$, $RMSEA = 0.067$.

4.3.2 Main Study

The questionnaire was sent to users from the iPhone thread ranking the highest on betweenness centrality, pagerank, outdegree centrality, specification-attributes-in-time and authority metrics. The highest 200 ranked people on each metric were sampled. That resulted in a set of 383 users as these groups overlap. Furthermore, the questionnaire was sent to 281 randomly selected users from top 0.05% (5707 users) the most active users in terms of outdegree centrality. The random group does not overlap with the other groups. In total, 664 users were approached.

The users were sampled in this way, because it was intended to make sure that there is enough lead users participating. The expectation to sample more lead users from the top scoring users stems from the fact that lead users tend to score higher on some of network metrics. In contrast, random sampling would involve mainly ordinary users.

There were 58 users who participated in the survey. The response rate is 8.7%. In fact, the users were contacted in batches. In the first batch, the highest top 100 ranked users on each metrics and around 100 random users were contacted. From the first batch, 56 people participated. In the second batch, the top 100–200 ranked users and around 200 random persons were contacted. From the second batch, only 2 users responded.

The answers with no user name provided were dropped, because they cannot be used as labels; six answers were without a user name. There were no other missing items in the responses. In the next step, the reliability was inspected. Std. Cronbach’s alpha of all the constructs except of CBR were above the acceptable level of 0.7 (see Table). CBR was therefore dropped. That left five constructs measuring lead userness. Item-total correlations of AOT3 is 0.67, PRK2 is 0.65 and OL2 is 0.68. All other item-total correlations are higher than 0.7. Although three items do not reach 0.7, this is considered acceptable given the sample size. Confirmatory factor analysis suggests a poor fit for data, $p = 0.001$, $CFI = 0.82$, $TLI = 0.77$, $RMSEA = 0.11$. This is argued to occur as a cause of the limited sample size as well. Furthermore, 93.1% of the participants own an iPhone. The other smartphone brands are represented sparsely.
Table 7: Std. Cronbach’s Alpha of Measured Constructs in the Main Study.

<table>
<thead>
<tr>
<th>Construct</th>
<th>AOT</th>
<th>HRB</th>
<th>PRK</th>
<th>INV</th>
<th>OL</th>
<th>CBR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std. Cronbach’s Alpha</td>
<td>0.76</td>
<td>0.74</td>
<td>0.76</td>
<td>0.77</td>
<td>0.71</td>
<td>0.59</td>
</tr>
</tbody>
</table>

The questionnaire’s scale ranges from 15–75. Following the work of Belz and Baumbach [5], a lead user is defined as a user scoring in the top quantile (with a minimum score of 58.35) of the scale. That means that a user needs to score in average at least 3.89 out of 5 points per item to be considered a lead user. Based on this criteria, there were 32 lead and 20 non-lead users identified.

4.4 Results & Discussion

To answer the first research sub-question, it is examined whether the metrics contain at least 10 lead users among the top scoring users. The top 300 scoring users of each metric are plotted against the cumulative amount of lead users (see Figure 11). The figure shows that indegree centrality, hub, pagerank, authority, and betweenness centrality metrics contain at least 10 lead users (dashed line in the figure) among the top 100 ranked users on these metrics. That is a sufficient amount for a NPD project with a typical level of lead users involvement [64]. Additionally, votes, closeness centrality, outdegree centrality, and specification-attributes-in-time contain a set of 10 lead users among the top 150 ranked users and can help to identify lead users as well. Latent-attributes and domain-interest showed to be less useful, as they contain less than 10 lead users in the top 300 scoring users.

![Figure 11: Amount of Lead Users among the Top Scoring Users of the Analyzed Metrics.](image)

To answer the second research sub-question, it is examined by a correlation analysis what is the relationship of the metrics and lead userness (i.e., the score from the questionnaire). The correlation analysis shows that no metric is related with lead userness (see Table 8). This means that in average a group of lower ranked users might contain more lead users than a group of
the top scoring users. This result can be explained by a limited data sample and also by the survey’s sampling, in which mainly the top scoring users on several metrics were approached. T-test inspecting the difference between the lead user and the non-lead user groups did not reveal any significant relationship either.

Table 8: Correlations of the Metrics and Lead Userness.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Correlation</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outdegree centrality</td>
<td>-0.05</td>
<td>0.72</td>
</tr>
<tr>
<td>Indegree centrality</td>
<td>-0.89</td>
<td>0.54</td>
</tr>
<tr>
<td>Domain-interest</td>
<td>-0.16</td>
<td>0.26</td>
</tr>
<tr>
<td>Betweenness centrality</td>
<td>-0.02</td>
<td>0.86</td>
</tr>
<tr>
<td>Closeness centrality</td>
<td>0.18</td>
<td>0.20</td>
</tr>
<tr>
<td>Hub</td>
<td>-0.08</td>
<td>0.55</td>
</tr>
<tr>
<td>Authority</td>
<td>-0.09</td>
<td>0.54</td>
</tr>
<tr>
<td>Pagerank</td>
<td>-0.04</td>
<td>0.75</td>
</tr>
<tr>
<td>Latent-attributes</td>
<td>-0.06</td>
<td>0.67</td>
</tr>
<tr>
<td>Specification-attributes-in-time</td>
<td>-0.12</td>
<td>0.40</td>
</tr>
<tr>
<td>Votes</td>
<td>-0.04</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Furthermore, it is inspected whether specification-attributes-in-time metric does better than its baseline. Specification-attributes-in-time and the baseline appear close to each other (see Figure 12). In some occasions, the baseline is even better. That means that in the context of lead user identification, specification-attributes-in-time, in particular the proposed attribute weighting, does not add any value to a plain frequency count of specification attributes.

Figure 12: Specification-attributes-in-time Metric and its Baseline.

In conclusion, focusing on users scoring high on selected metrics and manually inspecting whether they are lead users can rapidly decrease the time of lead user identification, because
only 100 users need to be assessed to find a sufficient amount of lead users for involvement in NPD. To help domain experts with the examination, a decision support tool LUSA is designed and presented in the next section.

4.5 Lead User Search Assistant

Lead User Search Assistant (LUSA) is a tool that intends to help domain experts to manually assess the identified candidate lead users (i.e., the top scoring community users) in terms of their lead userness. LUSA provides an interface to identify latent attributes. Next to it, it enables to view community users from an aggregated perspective, in which the best scoring users and the rest scoring users of each metric are compared in terms of domain-interest score. This helps to select a metric which can best prioritize for the most interested users in the domain. Finally, LUSA presents the top scoring community users given a selected metric and provides a zoomed view, which helps domain experts with a manual assessment of the presented community users.

The main advantages of LUSA is that it helps to analyze not only users who actively participate in identification procedure, but all community users. LUSA is able to process data semi-automatically and present relevant information in a systemic way based on which a domain expert can judge a user’s lead userness. A domain expert therefore does not have to read all the messages in a community forum. If that would be the case with the iPhone thread, it would take 111.1 days to read all messages in the thread given an average reader’s speed of 200 words per minute [50]. Reading all messages of a community forum does not obviously occur in practice. User population is always sampled, so the identification takes less time, approximately 8.5 man-days [8]. Yet, this is still a significant amount of time. It is proposed that using LUSA can reduce this time to hours. Lastly, LUSA allows to modify most of the parameters used in the experiments, so new hypotheses can be explored.

LUSA is built as a website. It consists of three pages: Latent Attributes, Aggregated Lead User, and Zoomed Lead User.

4.5.1 Latent Attributes Page

In this page, latent attributes can be identified via Word2vec latent attributes extraction approaches using implicit and explicit latent contexts. The page contains of query, results and selected latent attributes components (see Figure 13). The components correspond to the headings in the figure.

Query component enables to specify necessary input for the latent attributes identification and perform a Word2vec similarity query.

Results component projects candidate terms for latent attributes provided by the similarity query. A domain expert can select any of the candidate terms considering them latent attributes. Additionally, a user can insert other additional attributes in the text-box below. It is important to emphasize that the process of querying and selecting relevant latent attributes is circular, as the identification process is exploratory. A domain expert is encouraged to use different parameters to extract various latent attributes.

Selected latent attributes component aggregates latent attributes selected throughout multi-
ple queries. When the latent attribute identification is done, defined specification attributes and extracted latent attributes are used as an input to compute text metrics in the following page.

4.5.2 Aggregated Lead User Page

Aggregated Lead User page, as displayed in Figure 14, shows the individual metrics with respect to an amount of users’ messages referring to smartphones (i.e., to domain-interest). The graph in Figure 14 shows the difference in scores of the best scoring users and the rest scoring users on each metric. This comparison is meant to exhibit which user group is the most involved in the smartphone domain. Sizes of the best and the rest scoring groups are adjustable by the progress bar displayed at the top of the figure. In the presenting setting, the best scoring group is composed from the top 50 scoring users. After a metric of interest is selected, the top scoring users of this metric are displayed in a zoomed view.

4.5.3 Zoomed Lead User Page

Zoomed Lead User page shows individual users in a zoomed view (see Figure 15). Users are presented and ranked with respect to a selected metric from the previous page.
For every user, there is a user-rank on a selected metric and a user name shown (see the top of the figure). Under a user name, there are five threads presented in which a user has been the most active in the past half year or at maximum of 500 posts. More threads can be seen after clicking on "»".

Below, there are displayed user-ranks on all of the analyzed metrics. Each metric shows a progress bar, a name, and a corresponding score. The progress bar visualizes the ranking with respect to other community users. It is shown in different colors. Green means that a user scores within the top 2%, orange that he or she scores with the top 2.01-5% and red means that a user scores worse than that. The ranges were selected arbitrary. A score of a metric refers to an absolute score. So for instance, the user displayed in the figure ranks 9th in terms of cumulative sum of votes with absolute score of 1660 votes.

Messages containing latent specification attributes are displayed below the rankings. For every message, the following information is presented: a number of days prior to an attribute’s release, an amount of votes, and sentences in which attributes are included. Attributes are highlighted by green color and font size. The bigger the attribute’s weight, the bigger the font size. The same overview is provided for latent attributes.

The last part of this page is a graph displaying specification-attributes-in-time and latent-attributes discussed by given a user. The points in the graph refer to messages containing attributes. The x axis reflects attributes’ release dates relatively to the time when these attributes were discussed. For instance, the point -50 refers to a message which was posted 50 days prior to an attribute’s release. The y axis designates an attribute’s weight, which corresponds to the weighting function of specification-attributes-in-time. Furthermore, the size of the points refers to a number of votes a message has received and color to a sentiment of a message (the more red, the more negative). The sentiment of messages is calculated based on an approach of Hutto and Gilbert [26]. These two information can help a domain expert to focus on messages which are
liked by many users and might contain interesting ideas and on messages in which community users might express their unfulfilled needs. Finally, when a user hovers on a message, the message text is shown. Due to the size issues, there is only 500 characters displayed, evenly distributed among attributes with the biggest weight.

As mentioned in the beginning, the graph also displays latent attributes. Latent attributes do not have a release date. It was arbitrarily chosen to appoint their release date at the time of -500, so the weight is the maximum of the scale.

5 Conclusion

The goal of this research was to find metrics which can help to identify lead users. It was defined that a metric helps to identify lead users if it contains a sufficient amount of lead users among the best scoring users on that metric. The sufficient amount was specified as 10 lead users, which is a satisfactory number to establish a NPD project with lead users involvement [64]. A metric is also considered to aid to find lead users if there is a positive relationship between a metric and lead userness score. Both criteria are mutually exclusive. Inspecting various metrics in relation with lead userness is beneficial, because different subgroups of lead users can be potentially identified. The research was performed as a case study from the smartphone domain based on data from a product-oriented community forum about iPhones.
The results showed that the top 100 scoring users of indegree centrality, hub, pagerank, authority, and betweenness centrality contain at least 10 lead users. That means that analyzing only the top 100 scoring users on these metrics is satisfactory to identify a sufficient amount of lead users. The size of population to be analyzed by domain experts can be therefore lowered from potentially all 113,030 community users to a manageable sample of 100 users. In practice, obviously not all community users are inspected. In screening, approximately 2000 community users need to be approached to identify around 20 lead users [8, 38]. The chance of sampling a lead user is therefore around \( \frac{1}{100} \). That is 10 times less comparing to a chance of \( \frac{1}{10} \) when the top 100 scoring users are examined. Furthermore, votes, closeness centrality, outdegree centrality and specification-attributes-in-time metrics can help to identify lead users as well. They contain at least 10 lead users among 150 the top scoring users. Latent-attributes and domain-interest appeared to be less useful.

With respect to the second research sub-question, the results showed that there is no significant relationship between the metrics and lead useriness. That means that a group of lower ranked users can consist of more lead users than a group of the top scoring users. This result can be explained by the limited data size or the sampling strategy of the screening procedure.

To comment on intermediate research results, two new approaches for extraction of latent attributes from unstructured text were introduced. They are built on word embeddings of Word2vec. They outperformed latent attributes extraction based on template matching. The template matching approach did not extract any latent attribute. In contrast, the new approaches based on Word2vec were able to extract 19 latent attributes in total. This can be interpreted that 19 new ideas how to expand an iPhone are provided. These ideas represent needs which certain community users possess prior to other people.

Noteworthy, specification-attributes-in-time is a newly defined metric. It uses weighted specification attributes by time. The attribute weighting reflects the model of innovation diffusion by giving a more weight to attributes discussed earlier. As a consequence, the weighting distinguishes attributes which are posted prior to their release (i.e., latent specification attributes). Since latent specification attribute can be considered trends, using this metric might enable to skip trend identification in LUM which can take about two months and cost up to $9,000 [24]. Specification-attributes-in-time was compared with a baseline, which omits the attribute weighting. It appeared that specification-attributes-in-time metric does not perform better than the baseline and therefore the attribute weighting is not useful for lead user identification.

Lastly, a tool LUSA to help domain experts with lead users identification was built. LUSA facilitates latent attribute identification. It also enables to rank users based on a selected metric and then it provides an interface for domain experts to manually assess the top scoring users in terms of their lead useriness. The main advantage of LUSA is that it analyzes all community users and presents ones with a higher potential to be lead users. It is expected that a domain expert can shorten lead user identification time from a usual duration of 8.5 man-days [8] to hours.
5.1 Main Contributions

This research contributed to the area of automated lead user identification. It showed new approaches how lead users can be identified in a short time. The approaches can lower the size of population to be examined by domain experts during lead user identification procedure. This can enhance existing identification techniques. In netnography for instance, the most active users are manually assessed [5]. Instead, more effective metrics can be used to sample community users. As a considerable contribution to the lead user theory, this research examined which metrics are more effective for lead user identification. In this research, there were examined metrics which have and have not been previously analyzed with relation to lead userness.

The previously examined metrics include indegree centrality, outdegree centrality, betweenness centrality, and latent-attributes. Relating these metrics to previous research, centrality degree as the combination of indegree and outdegree centrality was reported not to be related with lead userness [33]. This research found no significant relation either, however showed that both indegree and outdegree centrality can reduce user population to be examined when searching for lead users. In contrast, betweenness centrality was reported to be related with lead userness [33]. Although, this research did not confirm this finding, it showed that there is a sufficient amount of lead users among the top scoring users of betweenness centrality. Lastly, mainly anecdotal evidence was reported in previous research about using latent attributes for lead user identification [59]. This research indicated that using only latent attributes does not help to identify lead users, arguably because they are too sparsely represented in data.

The research also proposed using new metrics which have not been examined before in a connection with lead userness. These are closeness centrality, pagerank, hub, authority, domain-interest, specification-attributes-in-time, and votes metrics. The results showed that these metrics can be useful for lead user identification. Apart from domain-interest, the top scoring users of these metrics contain a sufficient amount of lead users for involvement in NPD.

Specification-attributes-in-time is a newly defined metric. Its newly proposed part, the attribute weighting, turned out not to be useful for lead user identification. Despite that, specification-attributes-in-time approaches lead user identification from a completely different perspective. It is perceived that it has a potential to complement other metrics and identify a different sub-group of lead users. This metric also shows a way how different adopter groups might be identified.

This research showed a more precise and time-efficient method of extraction of latent attributes. Since, latent attributes can be interpreted as unfulfilled needs, which certain community users possess prior to other people, the research also contributed to the area of market research analysis.

The tool LUSA presents a major practical contribution. It is the second proposed software tool (after Pajo [47]'s FLUID), which facilitates semi-automated lead user identification. Moreover, it helps to manually assess community users in terms of their lead userness and facilitate latent attribute identification.
5.2 Limitations & Future Research

This project has been done as a case study of smartphones. Analyzing different domain might result in different conclusions. It is expected that it would be more difficult to establish this research in a low tech and low involvement product domain. This is because such domains are less vividly discussed on social media, so less data would be available for the analysis. Moreover, it might be difficult to establish text metrics relaying on prevalence of product attributes, because products from these domains consist of less product attributes and product specifications change less often in general. This might be also an issue for product categories in which new products consist of only attributes improved comparatively to previous models.

The major limitation of this study is the sampling strategy of screening to obtain labels for the users’ lead userness. There were sampled 200 the top scoring users on betweenness centrality, pagerank, outdegree centrality, specification-attributes-in-time, authority metrics, and 300 randomly selected users from 0.05% of the top scoring users in terms of outdegree centrality. There are two issues. The first is that the approached users might not contain any of the top scoring users of metrics which were not used for sampling. The second issue is that to inspect whether there is a significant relationship between a metric and lead userness, lower ranked users should also be included. The sampling was done in the way like it is, because the research question varied by time and the initial research goal was different. It is critical for future research to correctly choose a sampling strategy and possibly also select a different approach to obtain labels. Using a technique like netnography, in which domain experts are involved, would be more reliable than screening [5]. Another advantage of netnography is that labels of all community users can be obtained. In contrast, screening resulted in only 52 valid labels. Furthermore, with domain experts involvement, it would be interesting to evaluate e.g. innovativeness of extracted latent attributes, lead userness of identified lead users or usefulness of LUSA.

The proposed approaches of latent attribute extraction do not provide a context in which candidate latent attributes appear. Thus, it is very difficult to judge whether a term is or can be a latent attribute. Future research should focus on providing this context. Moreover, it would be interesting to introduce a weight to latent attributes, because they are not equally important.

The parameters of text metrics were not explored in depth due to the time scope of the research. Future research could inspect for instance different weighting functions of specification-attributes-in-time metric or different context words of latent-attribute metric.

Specification-attributes-in-time metric is based on using specification attributes from a product category of interest. Future research could inspect using specification attributes from different product categories. For instance to examine whether users discussing about attributes of gaming devices in context of smartphones are more likely to be lead users in the smartphone domain. This idea originates from the fact that innovation often occurs as a combination of different areas [33].
References


Appendices

A Lead User Characteristics

General lead user characteristics are presented in Table 9 and Table 10. Network characteristics are in Table 11.
<table>
<thead>
<tr>
<th>Product Field</th>
<th>Method</th>
<th>Lead User Characteristics</th>
<th>Source</th>
</tr>
</thead>
</table>
*Sample*: 129, 193, 139 people (per field).  
*Data collection*: Surveying community members in corresponding product fields.  
*Lead userness measurement*: Seven, eight, and nine items constructs adapted from Franke et al. [16]. | - knowledgeable  
- have a use experience  
- high locus of control  
- innovative  
- fast adoption of new products | Schreier and Prugl [53]. |
*Sample*: 153 people.  
*Data collection*: Surveying customers of two randomly selected manufactures in Germany.  
*Lead userness measurement*: By asking participants whether they have innovated or have an innovative idea. | - have unfulfilled needs  
- are dissatisfied with existing products  
- enjoying innovating process  
- have a use experience  
- have product related knowledge  
- financial rewards is not a predictor | Lüthje [40]. |
*Sample*: 249 members with 1,735 posts.  
*Data collection*: Manual from a community forum.  
*Lead userness measurement*: Community users displaying at least five out of six lead user characteristics are lead users. | - have unfulfilled needs  
- are dissatisfied with existing products  
- enjoying innovating process  
- have a use experience  
- have product related knowledge  
- financial rewards is not a predictor  
- are opinion leaders | Belz and Baumbach [5]. |
<table>
<thead>
<tr>
<th>Product Field</th>
<th>Method</th>
<th>Lead User Characteristics</th>
<th>Source</th>
</tr>
</thead>
</table>
| Basketball shoe online consumer community. | Method: Netnography.  
Sample: 460 discussion threads including 11,000 posts.  
Data collection: Manual from five basketball communities counting 11,200 members.  
**Lead userness measurement:** Lead users are considered enthusiastic and creative basketball players with a use experience and product-related knowledge. | - excitement driven  
- play basketball  
- between 20-25 years old  
- long time and highly acknowledged community members  
- dream of becoming professional shoe designer  
- often study design or art  
- demonstrate extraordinary drawing skills  
- have a deep knowledge about current and past basketball shoes  
- usually share the same opinion about shoes design with other lead users  
- are users with the most innovative content, not the highest number of posts  
- freely share their ideas and insights  
- gives advises to other designers | Jawecki and Füller [29]. |
Table 11: Network Characteristics of Lead Users.

<table>
<thead>
<tr>
<th>Product Field</th>
<th>Method</th>
<th>Lead User Characteristics</th>
<th>Source</th>
</tr>
</thead>
</table>
| School        | Method: Survey - questionnaire.  
Sample: 267 people.  
Data collection: Surveying students.  
Lead userness measurement: Two constructs adapted from Franke et al. [16], Morrison et al. [45], and Morrison [44]. | - higher degree of betweenness centrality  
- centrality degree is not a predictor | Kratzer et al. [33]. |
| Multi-industrial | Method: Secondary Research.  
Sample: 3118 people.  
Data collection: Manually from interviews with customers of 11 companies.  
Lead userness measurement: Provided by the companies. | - higher degree of betweenness centrality  
- centrality degree is not a predictor | Kratzer et al. [33]. |
| Aerospace     | Method: Netnography.  
Sample: 50 users with the highest degree of betweenness and 150 randomly selected users.  
Data collection: Automatized from an aerospace community space forum counting 431,257 posts from 13,287 individuals.  
Lead userness measurement: Two constructs adapted from Belz and Baumbach [5]. | - higher degree of betweenness centrality  
- centrality degree is not a predictor | Kratzer et al. [33]. |
B Literature Review about Relation Extraction

Relation extraction is one of the sub-tasks of information extraction. Its goal is to extract relations between entities from an unstructured text. For instance, organizations and their location relations can be a subject of relation extraction. As only two entities are present in the relation, it is a binary relation. This section further focuses on a review of binary relation extraction, because product attributes are typically extracted in this way (e.g., [59]).

There are three possible ways, how to address binary relation extraction and this is by supervised, unsupervised, and semi-supervised learning [31]. Each approach is further described in the following subsections. Note, that this section is mainly based on the materials introduced in a class given by Jurafsky and Manning [31] in Stanford University.

B.1 Supervised Learning Approach

In general, methods based on supervised learning perform well in binary relation extraction. A disadvantage is that they are domain dependent [43] and they need labeled data. Since no such data are available in this project, the review of supervised learning methods is skipped.

B.2 Semi-supervised Learning Approach

Semi-supervised relation extraction is typically done via bootstrapping or distant supervision.

B.2.1 Bootstrapping

Bootstrapping is a relation extraction method which needs only a few labeled seeds to be initialized with [22]. Then, it iteratively learns patterns based on provided seed examples and uses these patterns to extract new relationship values. It is important to emphasize that bootstrapping is based on data redundancy. It cannot extract every possible value, but only those appearing in structures being present in text multiple times [22].

Bootstrapping was first applied by Brin [9] in a system called DIPRE to capture relations between authors and books they had written. A pattern in DIPRE is defined as a tuple of prefix, middle and suffix [9]. Prefix is composed from \( m \) characters preceding an author (or a title if it is mentioned first). Middle part refers to a string between an author and a title [9]. Suffix is composed of \( m \) characters after a title (or an author).

DIPRE extracts a new author-book pair from a sentence if this sentence matches one of the earlier extracted patterns. In every iteration, all newly extracted pairs of authors and books are grouped by their middle part and the longest common suffix and prefix are determined. Based on that, new patterns are created. To prevent having too general patterns (e.g., " " from a sentence “Shakespeare Hamlet.”), a pattern specificity is introduced. It drops patterns that are too general.

Brin [9] ran an experiment with a seed of five relations of titles and authors. DIPRE was able to extract 15,000 new relation pairs from approximately 5 million web pages [9]. A manual inspection of 20 randomly selected titles and corresponding authors showed that 19 out of 20 relations were identified correctly.
Agichtein and Gravano [2] introduced a relation extraction system called Snowball, which is used to extract location-organization pairs. Snowball includes entity tagging as one of two major improvements of DIPRE. That has an important effect on precision of extraction. Snowball given a sentence “the Armonk-based IBM has introduced.” and a seed pair (Armonk, IBM) extracts a pattern <location>-based<organization>7. Unlike in DIPRE, this pattern would not match sentences like “computer-based learning”. The second improvement concerns matching between a pattern and a candidate sentence using confidence intervals as a measure of patterns’ trustworthiness. Untrustworthy patterns are filtered based on a proportion of newly extracted pairs disagreeing with previously extracted pairs. In general, a new organization-location pair is identified if it matches at least a few trustworthy patterns or more untrustworthy ones. Due to these advancements, Snowball turns to perform better than DIPRE [2].

B.2.2 Distant Supervision

Distant supervision combines bootstrapping with supervised learning [31]. It uses a big amount of known seed examples to create a feature space, which is further facilitated to employ a supervised classifier [31]. It can be used to extract hypernym relations from text. A hypernym relation represents an is-a binary relation. An entity X is-a hypernym of Y if Y is a sub-type of X [56]. For instance, a term city is a hypernym of a term New York.

Distant supervision to extract hypernym relations was used by Snow et al. [56]. Their approach goes as follow. Known noun hypernym pairs from semantic database WordNet are extracted to establish a set of seed examples. The known non-hypernyms pairs are extracted as well to feed a classifier with negative examples. Subsequently, a data corpus of 6 million news sentences is parsed and dependency paths (i.e., lexicon-syntactic patterns) between known and hypernym and known non-hypernym pairs are recorded. Every link on a dependency path between two words is represented by a stem of the words, their POS tags, and syntactic relations between the words. In the following step, extracted paths serve as features to build logistic regression and Naive Bayes classifiers to determine whether an unseen relation pair can be categorized as a hypernym relation.

Mintz et al. [43] also extracted relations from text in similar way. They followed up on Snow et al. [56] and additionally applied word chunking, in which consecutive words with the same entity tags are merged together and treated as one. They used Freebase semantic relations as seeds and data corpus of 1.8 million sample pages from Wikipedia. It was examined and showed that using both lexicon and syntactic patterns yields better performance than using them individually. There were 10,000 instances of Freebase relations extracted at precision rate of 67.7%.

B.3 Unsupervised Learning Approach

Two presented unsupervised learning approaches are TextRunner and Vector Space Models.

7The suffix and prefix is omitted just for illustrative purposes.
B.3.1 TextRunner

TextRunner was built to address shortcomings of relation extraction approaches using de-
pending parsing [3]. These approaches run into problems when they are used for heterogeneous
text on the Internet and when they have to deal with speed issues [3].

TextRunner consists of self-supervised learner, single-pass extractor, and redundancy-based
assessor. Self-supervised learner given a small corpus sample provides a classifier that deter-
mines whether a candidate relation pair is trustworthy or not. A candidate pair is labeled as
positive, when all pre-defined rules about a sentence are met. Some presented rules are [3]:

- A dependency path between a relation pair is no longer than a fixed length.
- A dependency chain between a relation pair does not cross sentence boundary.
- Either of entities in a relation do not solely consist of a pronoun.

Note that, dependency parsing is actually used to train the extractor, but it is not used for the
classification.

In the following step, features including e.g. POS tag sequences between pair nouns, the
number of stop words, and the number of tokens are recorded and labeled as trustworthy if
fulfilling the criteria. These features are profound to be domain independent [3].

Then, single-pass extractor POS tags all the words, remove non-essential phrases such
prepositional phrases, and identify noun-phrases. Finally, using Naive Bayes classifier it is
determined if a candidate relation pair is trustworthy (i.e., is a valid relation).

TextRunner was compared with at the time state-of-art extraction system KnowItAll [12].
It was found that recall decreases by 1.4%, however average extraction error rate improved by
33%.

B.3.2 Vector Space Models

Vector space models (VSMs) represent, or in other words embed, terms in a continuous
vector space. VSMs can be categorized into count-based (e.g., latent semantic analysis) and
predictive methods (e.g., neural probabilistic language models).

Count-based methods predict a term based on a collocation frequency of its neighboring
terms. Predictive models predict a term by more complex techniques. Neural networks can
be for instance used for the prediction. Predictive models outperform count-based methods in
terms of typical semantic tasks like semantic relatedness or synonym detection [4]. For this
reason, predictive models are further inspected.

Arguably, the most popular predictive VSM is Word2vec, introduced by Mikolov and Dean
[42]. Word2vec is a computationally-efficient model for learning word embeddings from text.
It learns word embeddings either by Continuous Bag-of-Words model (CBOW) or Skip-Gram
model. The difference in between them is that CBOW predicts target terms (e.g., "work") from
source context words (“... is going to work in the afternoon”) and Skip-Gram does the opposite
and predicts source context words from target terms. In general, CBOW is more useful for
smaller datasets and Skip-Gram for larger datasets. In practice, there are many use cases where
Word2vec can be applied.

---

8This section is written based on Tensorflow [57].
C Automated Extraction of Product Specification Attributes

There are several options how to identify a reliable source from which product specifications and further specification attributes can be retrieved. They include using APIs from Google Product, CNET, Semantics3, Amazon or developing a manual crawler. Given the emphasis on maximal generality of the solution and the implementation time, APIs of Amazon and Semantics3 were further considered to be used.

C.1 Data Sources

C.1.1 Amazon API

Amazon is the "Everything Store". It sells a gigantic variety of products. Although, Amazon API is easy to use, it does not allow to query products per product category. That means that to retrieve attributes for a given product category, it would be necessary to specify product names beforehand and also to make sure that these names would lead to correct pages on Amazon. An another issue with Amazon is that it does not guarantee how extensively are the product attributes specified. Observing several products and retrieving their attributes led to very limited results. The first listed product under query "iphone" to the date of 20. 4. 2016 has only four attributes specified. Amazon API therefore seems not to be a suitable source for extracting product attributes.

C.1.2 Semantics3 API

Semantics3 is a specialized service focusing on providing meta data about products including for instance product prices, images, URLs, and also product attributes [54]. Semantics3 provides access to more than 10,000 product categories. As a big advantage, its API offers a "per category" search. Semantics3 seems to fit the project requirements well. It is therefore further evaluated whether provided product specifications are relevant and complete.

C.2 Evaluation

To inspect whether retrieved attributes from Semantics3 are relevant (i.e., whether they belong to products from a selected category) and are suitable to be searched on social media, I examined results from six arbitrary selected product categories in terms of precision. The categories include mobile phones, books about cars, electric shavers, desktop computers, sofa and chairs, and soccer equipment categories. The mobile phone category is also evaluated in terms of recall. The benchmark set to evaluate recall is an aggregated set of attributes of 10 phones with the most fans according to GSMarena [21]. Let denote this set as \( A_{gsm} \). This set is not a complete set of phone attributes. However, the most important ones are present, assuming that GSMarena [21] would not skip them as it is a popular site with product specifications.

To measure aforementioned evaluation metrics, information including a total amount of attributes retrieved \( A_{all} \), an amount of attributes composed from one or two words \( A_{short} \), an amount of irrelevant attributes \( A_{short-ir} \) and an amount of not suitable attributes \( A_{short-ns} \) were assembled. Sets \( A_{short-ir} \) and \( A_{short-ns} \) are subsets of \( A_{short} \). \( A_{short-ns} \) is a set of attributes...
containing parenthesis (e.g., dimensions (wXhXd)), a logical or (e.g., internet/email capable) or a verb (e.g., is portable). There is a little chance to find these attributes on social media thanks to their unique notation. $A_{short}$ set was captured, because three and more words attributes also have a unique notation and are not expected to be useful. So that, set $A_{short}$ is evaluated instead of $A_{all}$.

Let denote relevant attributes as $A_{relevant} = A_{short} - A_{short-ir} - A_{short-ns}$. Precision can be further defined as $|A_{relevant}| / |A_{short}|$ and recall as $|A_{relevant} \cap A_{gsm}| / |A_{gsm}|$. The labeled sets and the evaluation metrics are stated in Table 12.

<table>
<thead>
<tr>
<th></th>
<th>Mobile Phones</th>
<th>Books about Cars</th>
<th>Electric Shavers</th>
<th>Desktop Computers</th>
<th>Sofa and Chairs</th>
<th>Soccer Equipment</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>A_{all}</td>
<td>$</td>
<td>304</td>
<td>39</td>
<td>135</td>
<td>446</td>
</tr>
<tr>
<td>$</td>
<td>A_{short}</td>
<td>$</td>
<td>204</td>
<td>32</td>
<td>94</td>
<td>251</td>
</tr>
<tr>
<td>$</td>
<td>A_{short-ir}</td>
<td>$</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>$</td>
<td>A_{short-ns}</td>
<td>$</td>
<td>7</td>
<td>0</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>Precision</td>
<td>0.96</td>
<td>0.84</td>
<td>0.89</td>
<td>0.94</td>
<td>0.73</td>
<td>0.25</td>
</tr>
<tr>
<td>Recall</td>
<td>0.37</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

It can be concluded that Semantics3 provides relevant product attributes. Only the category of soccer equipment has a precision lower than 70%. The recall of smartphone attributes is only 36.59%.

One more aspect to deal with is the ambiguity of the extracted attributes. For example, one of the attributes of the mobile phone category is an attribute band. In the domain of mobile phones, it refers to a cellular frequency of a phone. However, this term can also stand for a music band. This problem exists regardless a selected data source and will require manual effort to address it.

D Domain Information about Smartphones

Attributes of iPhones models with respective release dates are presented in Table 13. Note, that only new attributes comparatively to the previous models are stated in the table. The attributes were also used in their plural tense in the experiments.

Product names used for template matching are presented in Table 14. Other brands than Apple are in the table, because they are discussed in the data corpus too.
Table 13: Attributes of iPhone Models and their Release Dates.

<table>
<thead>
<tr>
<th>iPhone</th>
<th>Release</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2007-06-29</td>
<td>wlan, wifi, battery, ram, random access memory, memory, bluetooth, speaker, loudspeaker, usb, camera, gpu, sim, lcd, resolution, cpu, chipset, chipsets, os, operating system, display, radio, gsm, wcdma, cdma, card slot, keyboard, arm, proximity sensor, light sensor, microphone, multitouch, digital zoom</td>
</tr>
<tr>
<td>3G</td>
<td>2008-07-11</td>
<td>3g, gps geotagging, autofocus, auto focus, voice dialing, cortex a8,</td>
</tr>
<tr>
<td>3GS</td>
<td>2009-06-19</td>
<td>oleophobic display, hsdpa, digital compass, compass, voiceover, voice control</td>
</tr>
<tr>
<td>4</td>
<td>2010-06-10</td>
<td>retina, front camera, front facing, 720p, gyroscope, gyro accelerometer, motion coprocessor, facetime, digital compass</td>
</tr>
<tr>
<td>4s</td>
<td>2011-10-14</td>
<td>1080p, dual core, dualcore, speech recognition, voice recognition, siri, cortex a9</td>
</tr>
<tr>
<td>5</td>
<td>2012-09-21</td>
<td>nano sim, bottom mounted jack, hspa, lightning, lightning adapter, lightning connector, armv7</td>
</tr>
<tr>
<td>5s, 5c</td>
<td>2013-09-20</td>
<td>quad core, fingerprint, touch id, burst mode, dual led, image stabilization, ois, 64bit, apple a6, apple a7</td>
</tr>
<tr>
<td>6, 6 Plus</td>
<td>2014-09-19</td>
<td>video stabilization, nfc, nfc payment, mobile payment, barometer, dual domain, apple a8</td>
</tr>
<tr>
<td>6s, 6s Plus</td>
<td>2015-09-25</td>
<td>3d touch, optical stabilization, apple a9</td>
</tr>
<tr>
<td>SE</td>
<td>2016-03-31</td>
<td>-</td>
</tr>
</tbody>
</table>

*The iPhones release dates were retrieved from individual pages about iPhone models on Wikipedia.

E Statistics about Specification Attributes

The frequency counts of the latent and the feedback specification attributes are displayed in Figure 16 and Figure 17 respectively. Note, that the feedback specification attributes with frequency lower than 500 are not displayed in Figure 17 due to readability issues.

F Identified Latent Attributes

The latent attributes used in latent-attributes metric are displayed in Table 15. The attributes were also used in their plural phases in the experiments. The frequency count of the latent attributes are presented in Figure 18.

G Questionnaire

The questionnaire including the details about the reliability measures is displayed in Table 16.
### Table 14: Models of Smartphones.

| iPhone 3, iPhone 4, iPhone 4S, iPhone 5, iPhone 5S, iPhone 6, iPhone 6S, iPhone 6S Plus, iPhone 7, iPhone SE, iPhone, Samsung Galaxy, Samsung Galaxy 2, Samsung Galaxy, Samsung Galaxy 4, Samsung Galaxy 5, Samsung Galaxy Note5, Samsung Galaxy 6, Samsung Galaxy S6 Edge, Samsung Galaxy S6, Samsung Galaxy 7 Edge, Samsung Galaxy J7, Samsung Galaxy Grand, Samsung Galaxy Grand Prime, Samsung Galaxy Core, Samsung Galaxy Core Prime, Samsung Galaxy Note, HTC One, HTC One M9, HTC One A9, HTC 10, HTC Desire, LG V10, BlackBerry Priv, BlackBerry, LG G, LG G Stylo, LG K10, LG K7, LG G4, LG Leon, LG Tribute, LG Tribute 5, LG Venice, LG Volt, Nexus 5N, Nexus, Motorola Moto G4, Motorola Moto G4 Plus, Motorola Moto, Sony XperiaXA Ultra, Sony Experia, Xiaomi Mi Max, Xiomi, Xiaomi Redmi, Xiaomi Redmi Note, Xiaomi Redmi, Nokia Lumia 730, Nokia Lumia 725, Lumia 830, Lumia, Phone, Smart Phone. |

### Table 15: Latent Attributes Used in the Experiments.


### Figure 16: Frequency Count of the Latent Specification Attributes.
Figure 17: Frequency Count of the Feedback Specification Attributes.

Figure 18: Frequency Count of the Extracted Latent Attributes.
Table 16: Questionnaire Measuring Lead Userness.

<table>
<thead>
<tr>
<th>Construct</th>
<th>Items</th>
<th>Item-total Correlation</th>
<th>Std. Cronbach’s Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ahead of Trend</strong></td>
<td>1) In general, I am one of the first within my circles of friends who buys novelties in the area of smartphones.</td>
<td>0.94</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>2) I love buying novelties in the area of smartphones before the majority of people do.</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3) Generally, I belong to the first who uses smartphones products.</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td><strong>High Expected Benefits</strong></td>
<td>1) I often get irritated about the lack of sophistication in certain parts of smartphones.</td>
<td>0.76</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>2) I have needs related to smartphones which are NOT covered by the products currently offered in the market.</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3) I am dissatisfied with some functionality or design of available smartphones.</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td><strong>Product Related Knowledge</strong></td>
<td>1) I can repair my own smartphone.</td>
<td>0.82</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>2) I can help other people to solve their problems with their smartphones.</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3) I have difficulties making technical changes to my smartphone.</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td><strong>Involvement</strong></td>
<td>1) The advancement of smartphones matters to me.</td>
<td>0.86</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>2) The smartphone development is interesting to me.</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3) It is a lot of fun informing myself about smartphones.</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td><strong>Opinion Leadership</strong></td>
<td>1) I often influence people’s opinions about smartphones.</td>
<td>0.85</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>2) When others choose a new smartphone, they do NOT turn to me for an advice.</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3) I often help other people to buy a new smartphone that I like.</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td><strong>Community-based Resources</strong></td>
<td>1) If I wanted to make changes to my smartphone, I would know enough people who could help me do so.</td>
<td>0.86</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>2) I DON’T know many people, who have a thorough knowledge about smartphones.</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3) When I encounter technical problems with a smartphone, I know exactly who to ask for an advice.</td>
<td>0.84</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)This item was re-coded as 6 – score, because it is stated in negation.

\(^b\)The text of this item was adjusted in the main study due to the low clarity. The adjustments are mentioned in Section 4.3.1.