Automated Feature Extraction from Social Media for Systematic Lead User Identification

Sanjin Pajo, Dennis Vandevenne, Joost R. Duflou

Centre for Industrial Management, KU Leuven, Celestijnlaan 300 bus2422, Heverlee 3001, Belgium

{Sanjin.Pajo, Dennis.Vandevenne, Joost.Duflou}@kuleuven.be

Manufacturers strive to rapidly develop novel products and offer solutions that meet the emerging customer needs. The Lead User Method, emerging from studies on sources of innovation by the scientific community, offers a validated approach to identify users with innovation ideas to support rapid and successful new product development process. The approach has been more recently applied on online communities, where collection and analysis of rich user data are performed by expert practitioners. In this paper, feature extraction techniques are outlined, that enable automated classification and identification of lead users that are present in online communities. The authors describe two case studies to construct a classification model that is then used to identify online lead users for confectionery products, and to evaluate the outlined feature extraction techniques. The presented research points to opportunities in automated identification within the lead user approach that further reduce the resource and time costs.

Keywords: lead user identification; data mining; social networks; design management

1 Introduction

In a rapidly evolving marketplace, manufacturers strive to speedily uncover emerging customer needs, and to develop, and offer solutions that meet those needs. Defining future market needs is challenging to manufacturing teams, as explained by functional fixedness, where the understanding and evaluation of the product challenges are

bounded by the person's actual experiences in using a product (Allen & Marquis, 1964; Von Hippel, 1988; Chrysikou & Weisberg, 2005; Gavetti et al., 2005). Extensive research into consumer engagement in innovation activities has shown that a very small subgroup of customers called lead users, experiences needs ahead of the marketplace and stands to benefit greatly by finding solutions to meet those needs (Von Hippel, 1988; Churchill et al. 2009). Lead users can provide an insight into the emerging trends and they invest time and effort in uncovering beneficial and successful and commercially attractive solutions (Lüthje & Herstatt, 2004), with a significant portion of innovation accomplished by this subgroup of customers (Olson & Bakke, 2001; Schreier & Prügl, 2008).

As the product lifecycle is getting shorter and shorter (Guveritz, 1983; Pine, 1993; Dodgson, 2000), the producers have little time to identify and engage human resources like the lead users. The direction maintained in the established methodologies is employment of human resources to manually collect and analyse vast amounts of user data, resulting in significant costs in time and resources in identification of lead users. Advancements in data mining techniques signal opportunities towards minimizing resource and time costs, where identification of valuable human resources is automated and can be a matter of minutes instead of several weeks. Therefore, an automated systematic approach utilizing data mining techniques to identify online lead users is advocated by the authors. The two main phases in identification are: (1) gathering of the appropriate criteria or features for the purpose of the innovation project; and (2) screening of users and identifying lead users that meet these criteria (Bilgram et al., 2008). In this paper, the first phase is addressed, where the focus of the discussion is the set of techniques for automated extraction of user attributes or features from online data. To respond to the lack of automated and systematic user information retrieval for the

purpose of identifying online human resources at the fuzzy front end of the product design, the aim of this paper is twofold:

I. State and specify the techniques for extraction of user features through online media essential for the systematic identification of lead users.

II. Present case studies for a systematic approach utilizing the outlined feature extraction techniques.

The presented feature extraction techniques are meant to provide an example effective set for identification of online lead users.

Based on a literature study, summarized in Section 2, the existing techniques for extracting lead user data and identifying lead users are examined. Thereafter, in Section 3, a brief overview of the automated approach advocated by the authors is presented. A component of that approach, the extraction of the user features, and the entailed techniques are outlined. In Section 4, the execution and results of three case studies using the automated approach and the embedded feature extraction techniques are presented. The implications of the findings are discussed in Section 5, with the conclusions and outlook given in Section 6.

2 Literature Overview

Screening of a large number of potentially relevant users is the main method of lead user evaluation and identification (Belz & Bumbach, 2010). This offline approach is referred to as the Lead User Method (LUM). In the LUM approach, the screened individuals are evaluated utilizing survey questions or items measuring for validated lead user characteristics and engagement in innovation. Due to large population sizes and scarcity of lead users within the population, surveys or questionnaires are oftentimes not the most efficient approach to user information extraction and analysis (Belz & Baumbach, 2010). The smaller the number of lead users in a population, the

lower the efficiency and the higher the search costs (Sudman, 1985). In a study performed by Lüthje (2000), out of 2000 persons screened, 22 were identified as lead users, which is a low sample efficiency of 1.1%. The effectiveness studies indicate significant opportunities in increasing the efficiency of customer information extraction (Belz & Baumbach, 2010; Pollok et al., 2014).

To address the shortcomings of the screening approaches, like the low sample efficiency and the high search costs, more recently researchers have looked to web based lead user identification approaches. With the rapid development of Web 2.0 applications, the perception is that monitoring online communities or weblogs could replace the survey based approach to collecting customer information (Bilgram et al. 2008). Netnography, a fusion of Internet and ethnography concepts, is one such web based approach to systematically collecting and analysing data from online communities (Kozinets, 1998, 1999). In the Netnography approach, validated characteristics that allow for differentiation between lead users and other users in an offline context are extrapolated to the online context. Data collection and analysis entail direct copying of online user posts or exchanges, and the data researchers generate through observation of the users and their online behaviour. Belz and Baumbach (2009) demonstrate through the explorative study of the online community 'utopia' that a systematic online approach is a viable method of lead user identification. The identification efficiency was found to be greater than the efficiency of the mass screening searches. The researchers stipulate that for an experienced researcher it requires approximately 2-4 weeks to conduct an in-depth Netnography, still a significant amount of time that largely depends on the size of the online community.

Formulation of a process for effective lead user identification remains a challenge to the researchers in the field. Combined with the growth of collaborative

Web 2.0 based platforms like the social media and the reductions in communication costs, fast, systematic and automated analysis techniques can be utilized and evaluated towards identification of lead users, reducing the challenges faced by manufacturers to quickly deliver successful and radical solutions into the marketplace. In the following section, the automated and systematic Fast Lead User Identification (FLUID) approach for identification of lead users on social media, as developed by the authors in response to the perceived need for more resource efficient lead user identification, is briefly summarized.

3 FLUID approach

3.1 Methodology

The FLUID approach makes use of information retrieval, text mining, network theory and machine learning techniques to automatically collect and analyse user online information. The focus of this paper is on one of the steps in the process; therefore, the approach is briefly summarized here based on the previously reported research by Pajo et al., 2015. The main steps of the FLUID approach are: (1) scope definition, (2) medium selection, (3) automated lead user identification and (4) lead user engagement, as shown in the Figure 1 below.

Figure 1 FLUID phases.

In the first phase, the stakeholders, i.e. the company R&D team, decide on the domain or the product for which to identify lead users. Similarly to the Lead User Method, together with the company stakeholders the objectives, the utility the expected outcomes of the research are defined. Thereafter, online communities that are suitable for the chosen domain or the product are selected by the stakeholders. Next, automated data retrieval and analysis is performed to identify lead users present online. The step consists of user data retrieval, storage, feature extraction, and the classification of users based on the extracted features. In an online and machine learning context, the retrieved users are classified based on the empirically validated lead user characteristics that are transposed into online user features. Finally, the identified lead users are evaluated by the stakeholders for a possible engagement in NPD processes. In the following section, the feature extraction techniques employed for the FLUID automated lead user identification are outlined.

3.2 Lead User Feature Extraction Techniques

In the third phase of the FLUID approach to analyse the collected user data and make predictions, significant and relevant features or attributes are extracted from the data. These features should provide a characterization of users and their behaviour on a social media site and in consequence, allow for a clear differentiation between lead users and other users. The features to be extracted are based on the available metadata and the observed and validated lead user characteristics reported by Bilgram et al., 2008. In the following subsections, the extracted features and how they stem from the validated lead user characteristics are described in detail. The features can be logically classified into four distinct constructs: network centrality, activity, sentiment, and relevance measures.

3.2.1 Centrality Measures

One of the characteristics of lead users is the *opinion leadership*, the ability to facilitate the flow of information and in particular, diffuse the information, i.e. needs and solutions into the marketplace. A strong social relationship and engagement are necessary for a functioning exchange of ideas or knowledge transfer (Martínez-Torres, 2014; Arenas-Márquez et al., 2014). In 68% of the cases, as compiled by Bilgram et al. (2008), innovators have collaborated with two or more partners during the development of a new product, with teams of six present in 21% of the cases (Shah, 2000; Franke & Shah, 2003; Kozinets, 2006). Guided by network theory, Kratzer and Lettl (2008) showed that individuals who are positioned as bridging links between different groups in social networks reveal a high level of creativity. There is a strong link between facilitating flow of information and creativeness. In the following subsections, three fundamental indicators in facilitation of information flow are described and advocated for characterizing the position of a user in the social network: degree centrality, betweenness centrality and closeness centrality (Scott, 2000; Latora & Marchiori, 2007).

Degree Centrality

The degree centrality is the number of direct connections for a user to other users in the network. In the case of a social network, the graph network is generated by querying available relationships for each retrieved user. An example for a micro-blogging site is the retrieval of friend and follower ids for each user. Users are represented as nodes and directional edges are drawn depending on the type of friendship, follower or following. The degree centrality C_D for a vertex v, for a given network graph G is defined as (Newman, 2010):

$$C_D(v) = \deg(v) \quad (1)$$

The higher the degree of centrality, the greater that user's reach. There are also more opportunities for information absorption and exchange, broadening that user's product or *domain related knowledge*, which has been positively correlated to lead userness in previous studies (Lüthje, 2000; Schreier & Prügl, 2008).

Betweenness Centrality

To measure the ability of a user in facilitating information flow within the network, the betweenness centrality measure is calculated for each user in the constructed social network. The betweenness centrality C_B for a vertex v, for a given network graph G is defined as (Anthonisse, 1971; Freeman, 1977):

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (2)$$

In Equation 2, σ_{st} is the number of shortest paths from vertex s to vertex t and $\sigma_{st}(v)$ is the number of those paths that pass through vertex v. Individuals with higher betweenness centrality are able to connect disparate groups of users within the network and they can help diffuse information or solutions throughout the entire social network (Kratzer & Lettl, 2008; Kratzer et al., 2015).

Closeness Centrality

The closeness centrality is the average social distance between a vertex and every other vertex in a network graph. The closeness centrality C_C for a vertex v, for a given network graph $G := (V, E)$ where $|V|$ are vertices and $|E|$ are edges, is defined as (Bavelas, 1950; Sabidussi, 1966):

$$C_c(v) = \frac{1}{\sum_{t \in V} d_G(v,t)} \quad (3)$$

In the above Equation 3, $d_G(v,t)$ is the distance between the vertices v and t, i.e. the minimum length of the path connecting v and t in G (Sabidussi, 1966; Brandes, 2001). Nodes that are at the centre of a cluster tend to have higher closeness centrality values, meaning that individuals with high closeness centrality tend to be influencers in their particular communities or clusters (Chen et al., 2012).

Additional centrality measures that have been used in effectiveness studies, to give a more comprehensive depiction of the underlying network and the position of a user in the network, are: eigenvector centrality, farness centrality, hits centrality, eccentricity and page rank (Leskovec & Krevl, 2014). Their suitability may depend on the selected medium and the type of network data that is retrievable.

3.2.2 Activity Measures

The objective behind extracting activity measures is to depict user's behavioural profile, which entails the type and volume of activity. The presumption, based on the existing research is that the activity differs between various customer groups, for example, product early adopters on the one hand and so-called laggards on the other (Bilgram et al., 2008). In an online community, this can be the number of online posts over a specified period of time or the number of users befriended in the network per select unit of time. For each possible user action, the following is calculated:

$$y_a = \frac{|x_a|}{t - t_0} \tag{4}$$

In Equation 4, y_a is the activity a measure per time period, i.e. the number of posts, equal to the volume of the activity $|x_a|$ over a select period of time, in this case the difference in days from metadata retrieval and account creation, t_0. In the example of the micro-blog, scores can be calculated for the number of posts, followers and following added, number of lists and favourite counts. Additionally, participation, i.e.

frequency of contributions, and commitment to a social network are some of the indicators of *intrinsic motivation* (Bunz, 2006; Hemetsberger, 2001; Franke & Shah, 2003; Füller et al., 2007). They are an expression of feelings of enjoyment, relatedness, exploration and creativity that have been positively correlated with lead userness (Lüthje, 2000; Lakhani, 2006; Jeppesen & Frederiksen, 2006).

3.2.3 Relevance

Extensive knowledge of the product and product *use experience* are other principal characteristics of lead users. Both have been positively correlated to engagement in innovation (Füller et al., 2006; Schreier & Prügl, 2008). User discussions and posts can be ample indicators of awareness and knowledge in a particular domain, so therefore a measure for relevance of user's posts is extracted from the data. The relevance of user's posts is measured in respect to a set of keywords, terms selected by the stakeholders due to their significant business value and target domain clarification. For the FLUID test platform effectiveness analysis, a numerical statistic named term-frequency-inverse document frequency or tf-idf (Spärck Jones, 1972; Rajaraman & Ullman, 2011) is used to measure the relevance of a post $score(Q_t, p)$ to the collection of stakeholder keywords. The tf-idf score is computed as follows:

$$post_{relevance} = score(Q_t, p) = \sum_{(t)\in Q_t, p} tf(t,p) \cdot ln \frac{|P|}{df(t,P)} \qquad (5)$$

In the above Equation 5, Q_t is the set of the selected stakeholder terms, $|P|$ is the size of the post collection P, $tf(t,p)$ is the number of times the term t appears in the post p and $df(t,P)$ is the number of posts in P that contain the term t. User's timeline relevance score or the relevance score for the retrieved collection of user posts $score(Q_t, P_N)$ is computed as follows:

$$score(Q_t, P_N) = \frac{\sum_{p \in P_N}(score(Q_t, p))}{|P_N|} \quad (6)$$

The expectation is that there is a transfer of use experience between the lead users (Bilgram et al., 2008), communication of relevant information in purchasing, use and modification of a market solution.

3.2.4 Sentiment

Whether excitement-driven, where for users development of a product is an enjoyable activity, or driven by dissatisfaction, due to unmet needs, sentiment measures are a crucial gauge indicator towards differentiating between lead and non-lead users. Emotional disposition of user's posts is an indicator that can be measured. For example, the extraction process utilized for assessing the FLUID effectiveness makes use of a publicly available sentiment dictionary SentiWordNet (Esuli & Sebastiani, 2006) to calculate the overall sentiment expressed in online posts. In the lexicon, each synset or terms that are semantically equivalent, is assigned a triple score, i.e. positivity, negativity and neutrality score with a sum of these scores equal to 1 (Esuli & Sebastiani, 2006). To measure the sentiment, first the stop words are removed from user's post and each word in the post is stemmed. Then, each stemmed word is looked up in the SentiWordNet. Given n corresponding synsets for a word W, the following formulas are used to determine triple scores, $score_{pos}$, $score_{neg}$ and $score_{ntr}$ (Denecke, 2009; Esuli & Sebastiani, 2006):

$$score_{pos}(W) = \frac{1}{n}\sum_{i=1}^{n} score_{pos}\,(i) \quad (7a)$$

$$score_{neg}(W) = \frac{1}{n}\sum_{i=1}^{n} score_{neg}\,(i) \quad (7b)$$

$$score_{ntr}(W) = \frac{1}{n}\sum_{i=1}^{n} score_{ntr}\,(i) \quad (7c)$$

Thereafter, scores of all the terms in the post are added and divided by the number of terms in the post *n*. The results are positivity, negativity and neutrality values for each user's post. Finally, the percentages of neutral, positive and negative posts are calculated for each user, to be added as sentiment features.

The above outlined extracted features are not exclusive but allow for a distinctive depiction of user's behaviour and characteristics on social media and can be theoretically linked to the offline traits attributed to the lead users. The authors stipulate that the extracted activity, sentiment analysis, relevance and network measures are sufficient enough to describe user's behaviour online and to build a classification model that can be used effectively to predict lead userness. The significance of the extracted measures in the identification of lead users will be evaluated in the effectiveness analysis, Section 4.5. In the following section, the effectiveness of utilizing the above indexed features in an online community is described.

4 Case Studies

To evaluate the effectiveness of the systematic FLUID approach and in turn, the outlined feature extraction techniques, three case studies were executed. The first two test case studies, 'lens products' and 'scrap aluminium', were used to produce a set of training data to build a lead user classifier. The third industrial case, confectionary products, was used to validate the generated lead user classification model and the extracted features. The selection of the cases was done based on the availability of experts and industrial partners in a domain.

4.1 Social Network Selection

The micro-blogging site Twitter was selected as the preliminary test medium to evaluate the FLUID approach. It is a social network rich in data, with a large user base of over

300 million active accounts (Twitter, 2015). As an open network and with users with expertise in a variety of occupations, Twitter facilitates discussions on a wide spectrum of topics. The network affords sharing of multimedia content, i.e. text, image, video that can be effectively used for idea generation and sharing. It also provides an easy access Application Programming Interface (API) to retrieve structured user information, relevant posts or tweets and user metadata (Twitter Developers, 2015). More importantly, Twitter permits direct contact with users for future in-person interviews and users often include additional links and information on hobbies and professional background, as well as contact and location information in their profiles.

4.2 Data Collection

To find relevant discussions and users on the micro-blogging site Twitter, the FLUID test platform makes use of the Twitter search engine. For each test case, together with the experts in the field or the company innovation management team, a set of search terms has been defined to be used to collect relevant online data. As mentioned previously, terms were selected on the basis of the business value or significance they have to the stakeholders. The list of terms consists of bigrams and trigrams including product names, modifiers or parameters, to narrow down the search. Example bigrams for the confectionary products case are shown in Table 1 below.

Table 1 Example confections search terms.

Term A	Term B
Treat	Chocolate
Treat	Biscuit
Boost	Snack
Breakfast	Biscuit

To retrieve data from Twitter, the FLUID test platform iteratively requests relevant tweets and embedded user metadata through the Twitter search engine using the search

terms. For each identified user, related metadata were downloaded and friend and follower IDs were also retrieved for the ensuing centrality analysis. To alleviate strain on resources, Twitter imposes restrictions on the number of API calls, which constrain the data retrieval process to 15 minute intervals with at most 180 requests per interval (Twitter Developers, 2015). Historical data is limited to 6-9 days as of the request. Additionally, to ensure efficient and effective resource utilization; users with protected or private account were filtered and outlier users with more than 2000 connections were not stored in the database; Twitter users on average have 208 followers and follow 102 users (Beevolve, 2015). After a collecting data, interactive hashtag analysis was performed with the stakeholders to expand the search query and to increase the number of relevant posts and users retrieved. The first part of the process is the extraction of all the hashtags from the retrieved tweets and the selection of the most relevant and novel trending hashtags by the stakeholders to be added to the set of search terms. The second part is the selection of retrieved relevant clusters of users to be expanded, by retrieving friend and follower metadata. Table 2 below shows the amount of retrieved social network data for the three cases, taking into account FLUID test platform limits and Twitter imposed API limits and restriction.

Table 2 Retrieved Twitter users, messages and metadata.

Case	Users	Tweets	Friends and Followers	Directed Connections	Undirected Connections
Lenses	2,490	648,485	1,229,351	2,254,047	478,542
Scrap Aluminum	4,535	1,183,779	1,598,847	4,110,988	772,803
Confections	1,577	605,740	884,053	1,445,657	293,577

4.3 Feature Extraction

After user metadata were collected, centrality (Section 3.2.1), activity (Section 3.2.2), relevance (Section 3.2.3) and sentiment construct (Section 3.2.4) features were extracted for each case. For the activity construct, rates for the number of followers, friends, tweets, favourites, lists, etc. per day were calculated. For the centrality construct (Section 3.2.1), network centrality scores were calculated for betweenness, closeness, degree, eccentricity, eigenvector, farness, hubs, authorities, and page rank (Leskovec and Krevl, 2014). Tweet relevance was extracted by utilizing term frequency/inverse document frequency based on the query list and sentiment by making use of the SentiWordNet Lexicon (Esuli & Sebastiani, 2006). Additionally, the percent of relevant hashtags was extracted. In total, 25 features were extracted from the collected metadata for each retrieved user and all the values for each feature in the dataset were normalized in the range [0, 1] using the following formula:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (8)$$

Thereafter, a classification model was built using two of the datasets (Section 4.4) and tested on the data for confections (Section 4.5).

4.4 Classification Model

4.4.1 Training Data Generation

After data extraction and normalization, training data were aggregated using two of the cases: lens products and scrap aluminium. In a study by Pajo et al., 2014, the authors describe lead user identification through an online survey using validated questionnaire items that measure for characteristics of being a lead user. Metadata of lead users and non-lead users identified using the survey method were used as training data to construct the initial classification model. To gather additional training data, an expert

evaluation approach was used for the urban mining case. First, the classification model, built using the lens training data, was used to classify users. Thereafter, two experts on lead user identification evaluated the predicted lead and non-lead user social network data. The experts were provided with a link to the Twitter profile page and the available blog or personal site of each user. An inter-rater reliability was computed to measure the level of agreement between the experts using the Fleiss' Kappa statistic (Fleiss, 1971). The kappa, k is defined as:

$$k = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{9}$$

Here, $1 - \bar{P}_e$ is the max degree of agreement possible and $\bar{P} - \bar{P}_e$ is the actual degree of agreement for multiple raters. The strength of the agreement between the experts was found to be strong, with Fleiss' Kappa for 39 cases with two possible classes, lead user and non-lead user, providing a score of 0.840, with standard error (SE) equal to 0.087 and 95% confidence interval from 0.669 to 1. The percentage of observations for which experts were in agreement is 92.31%. The percentage of agreements expected by chance is 51.81%.

4.4.2 Classification Model Building

The lens and urban mining training data were combined to create a classification model to test and validate on an industrial case, for confectionary products. To create the classification model, cross validation was performed on the training data using a range of classification methods. Based on the cross-validation results given by the confusion matrix, two models were generated using Random Forrest (Breiman, 2001) and Sequential Minimal Optimization (SMO) algorithms in Weka (Hastie & Tibshirani, 1998; Keerthi et al., 2001; Platt, 1998; Hall et al., 2009). Based on the cross-validation, the selected Random Forrest model consists of 40 random trees where 5 random

features were used to generate each tree with non-limited depth (Breiman, 2001; Hall et al., 2009). The resulting confusion matrix, shown in Table 3 for 103 instances of training data (Twitter users), shows 94 or 91.26% correctly identified instances and 9 or 8.73% incorrectly identified instances with Kappa statistic 0.7819. The root mean squared error is 0.2706.

Table 3 Cross-validation confusion matrix (Random Forrest).

		FLUID Classifier	
		LU	**NLU**
True Values	**LU**	24	6
	NLU	3	70

The precision for the lead user class is 0.889 and recall is 0.800. The precision for the non-lead user class is 0.921 and recall 0.959. Additionally, the implementation of John Platt's SMO algorithm for training the support vector classifier was selected to build a model (Hall et al., 2009; Platt, 1998). The resulting confusion matrix, shown in Table 4, for 103 instances shows 86 or 83.50% correctly identified instances and 17 or 16.50% incorrectly identified instances with Kappa statistic 0.5796. The root mean squared error is 0.4063.

Table 4 Cross-validation confusion matrix (SMO).

		FLUID Classifier	
		LU	**NLU**
True Values	**LU**	19	11
	NLU	6	67

The precision for the lead user class is 0.760 and recall is 0.633. The precision for the non-lead user class is 0.859 and recall is 0.918.

4.5 Results

The two classification models, Random Forrest and SMO were used to classify 1577 retrieved users for the confectionary products case into lead and non-lead users. The

classification resulted in 23 lead users where both classifiers were in agreement, with the remaining 1554 users classified as non-lead users. To validate the results of the classified data set, the 23 predicted lead-users and randomly selected 23 non-lead users were evaluated by two experts. The experts were selected based on their expertise in the lead user method, their in-depth knowledge of validated lead user traits and experience in performing user evaluations. They were asked to indicate in a questionnaire form, if an observed Twitter user is a lead user or a non-lead user. To make this decision, each expert was provided with a link to a user's profile that included biographical information and the tweet timeline. Where available, links to other user profiles, for example blog sites, were also provided. Each expert was instructed to carefully examine a user's profile and read and evaluate user's posts by using several criteria before classifying that user as a lead or a non-lead user. The first criterion is the *ahead of trend*, meaning that the user is ahead of others in recognizing and planning new solutions to problems. The second criterion is the *high expected benefit*: a user can benefit significantly by the early adoption and use of newly developed products. The next two criteria, assessed user's engagement in innovative activities: user has ideas on how to improve products and makes improvements to the existing products. Finally, the last two criteria measured for *dissatisfaction* with the products offered in the marketplace: user has needs related to the products that are not covered by solutions offered in the market and user is constantly searching for improved products. The described criteria are validated traits that are used as a supplementary confirmation of lead userness (Bilgram et al, 2008) An inter-rater reliability was computed to measure the level of agreement between the experts. The strength of the agreement between the experts was found to be good, with Kappa for 46 cases with two possible classes, lead user and non-lead user, 0.698 and SE equal to 0.110. The percent of observations for which experts

were in agreement is 86.96%. The percent of agreements expected by chance is 56.81% of the observations. The level of agreement allowed for a comparison of the expert and FLUID platform classification results. The confusion matrix for the FLUID test platform results and expert evaluation results is shown in Table 5 below. Only the cases where both experts agreed on the classification of an instance are included in the confusion matrix.

Table 5 Experts vs. FLUID platform confusion matrix results for confectionary products.

		FLUID Classifier	
		LU	**NLU**
Expert Evaluation	**LU**	11	0
	NLU	8	21

The calculated overall accuracy for the confections case, based on the given confusion matrix is 0.80 with Kappa equal to 0.591. The calculated precision by which the FLUID platform predicted lead user is affirmed by experts to be a lead user is 0.579 and the calculated recall that a randomly confirmed lead user is retrieved by the test platform is 1. The accuracy of the model is substantial, although the results should be interpreted cautiously as not all the FLUID platform non-lead user cases were evaluated by the experts.

For evaluation of the features based on the training data, the Relief statistical selection algorithm was used to rank the constructs: centrality, activity relevance and sentiment (Kira & Rendell, 1992; Kononenko, 1994; Robnik-Sikonja & Kononenko, 1997, Hall et al., 2009). The algorithm has a relatively fast learning speed; it is noise-tolerant and selects only statistically relevant features (Kira & Rendell, 1992). Each feature is evaluated by repeatedly sampling an instance and considering the value of the given feature for the nearest instance of the same and different class. A feature is considered relevant when the relevance level is positive and irrelevant when the

relevance level is negative. The four highest ranked features in order were: tweet relevance $f_{trl} = 0.127$, betweenness centrality $f_{bc} = 0.119$, degree centrality $f_{dc} = 0.111$, and page rank $f_{pr} = 0.087$. Of all the features, only two, the tweet rate or the number of posts per day $f_{tr} = -0.01$ and the negative sentiment $f_{ns} = -0.009$, were found to be irrelevant. Thereafter, the average construct relevance levels were also calculated. All the constructs were found to have a positive average relevance level, with relevance, 0.08, and centrality, 0.07, constructs most relevant.

5 Discussion

The results of the expert evaluation for the urban mining and the confections cases indicate that the constructed statistical classifier is an effective model for accurately classifying lead users present on a social micro-blogging site. Additionally, the effectiveness analysis results signal that the extracted user features from the users' metadata, including centrality, activity, relevance and sentiment facilitate building an effective statistical classifier for online users. The extracted features are positively correlated with lead userness. Lack of high precision in the evaluation results necessitates an additional step of expert evaluation before the results are examined by the stakeholders. As is the case with other Netnography approaches (Belz & Baumbach, 2010), the FLUID method can be combined with targeted screening, improving the precision of the advocated approach. The sample efficiency is 0.7%, which is close to the sample efficiency of the lead user method approach, 1%, as reported by Lüthje (2000). The efficiency can possibly be improved by targeting online communities based around a single topic, as shown by Belz & Baumbach (2010).

The results must be interpreted cautiously as the effectiveness studies were performed on one micro-blogging site, Twitter, and for specific target domains. Results

might vary significantly depending on the producer's innovation project and the types of users targeted. The interpretations are also based on the online content and not on behaviour and observations of consumers in real life. Depending on the limitation of the selected network, additional indicators may need to be investigated for effective lead user identification. Nevertheless, the outlined techniques and the model constitute a functioning classifier for the micro-blog Twitter, and can be adapted to similar social networks with differing metadata structures and restrictions. The outlined extraction and classification techniques offer a nearly instantaneous analysis of collected online user data and are expected to be further refined through additional studies.

6 Conclusions

In this work the authors advocate a systematic approach, with automated retrieval and analysis of social network data using statistical techniques, to identify online lead users. As is the case with other Netnography approaches, the presented systematic FLUID approach further expands the Lead User Method from an offline to an online context. The opportunities provided by access to large amounts of structured social network data, as shown in effectiveness analysis examples can be more efficiently and effectively utilized by means of data mining and machine learning techniques. The approach therefore also advances the current web based Netnography approaches, by reducing the effort and time required to collect and analyse swaths of user online data. Finally, the advocated approach offers support in addressing the fuzzy-front end of new product development through rapid identification of human resources, lead users, to uncover emerging trends and needs, and partial solutions that address those needs. The presented work supports systematic processes towards aligning companies' activities with the emerging needs of the target consumers.

References:

Allen, T.J. and Marquis, D.G. 1964. Positive and Negative Biasing Sets: The Effects of Prior Experience on Research Performance. IEEE Transactions on Engineering Management EM-11 (4): 158-161.

Anthonisse, J. M. 1971. The rush in a directed graph. Technical Report BN 9/71, Stichting Mathematisch Centrum, Amsterdam.

Arenas-Márquez, F. J., Martínez-Torres, M.R. and Toral, S.L. 2014. Electronic word-of-mouth communities from the perspective of social network analysis, Technology Analysis & Strategic Management 26 (8): 927-942.

Bavelas, A. 1950. Communication patterns in task-oriented groups. J. Acoust. Soc. Am 22 (6): 725–730.

Beevolve. 2015. An Exhaustive Study Of Twitter Users Across The World - Social Media Analytics | Beevolve. [online] Available at: http://www.beevolve.com/twitter-statistics/ [Accessed 4 May 2015].

Belz, F.-M. and Baumbach, W. 2010. Netnography as a Method of Lead User Identification. Creativity and Innovation Management, 19 (3): 304-313.

Bilgram, V., Brem, A., & Voigt, K. 2008. User-centric innovations in new product development: Systematic identification of lead users harnessing interactive and collaborative online-tools. International Journal of Innovation Management 12 (3): 419-458.

Brandes, U. 2001. A faster algorithm for betweenness centrality. Journal of Mathematical Sociology 25 (2): 163–177.

Breiman, L. 2001. Random Forests. Machine Learning. 45 (1): 5-32.

Bunz, M. 2006. Wenn der Kunde handelt. Brand Eins 8 (4): 96-102.

Chen, D., Lü, L., Shang, M.-S., Zhang, Y.C., Zhou, T. 2012. Identifying influential nodes in complex networks Physica A 391 (4): 1777–1787

Chrysikou, E.G., and Weisberg, R.W. 2005. Fixation Effects of Pictorial Examples in a Design Problem-Solving Task. Journal of Experimental Psychology: Learning, Memory, and Cognition 31 (5): 1134-1148.

Churchill, J., von Hippel, E., and Sonnack, M. 2009. Lead User Project Handbook: A Practical Guide for Lead User Research Teams. Cambridge and Minneapolis: Lead User Concepts, Inc.

Denecke, K. 2009. Using sentiwordnet for multilingual sentiment analysis. In Data Engineering Workshop, 2008. ICDEW 2008, 507-512.

Dodgson, M. 2000. The Management of technological innovation. An international and strategic approach. (Oxford: Oxford University Press).

Esuli, A. and Sebastiani, F. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06), Genova, IT, 417–422.

Fleiss, J.L. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin 76: 378-382.

Franke, N. and Shah, S. 2003. How communities support innovative activities: an exploration of assistance and sharing among end-users. Research Policy 32 (1): 157-178.

Freeman, L. 1977. A set of measures of centrality based upon betweenness. Sociometry 40: 35–41.

Füller, J., Bartl, M., Ernst, H. and Mühlbacher, H. 2006. Community based innovation: How to integrate members of virtual communities into new product development. Electronic Commerce Research 6 (1): 57–73.

Füller, J., Jawecki, G. and Mühlbacher, H. 2007. Innovation creation by online basketball communities. Journal of Business Research, 60 (1), 60-71.

Gavetti, G., Levinthal, D.A., & Rivkin, J.W. 2005. Strategy Making in Novel and Complex Worlds. Strategic Management Journal 26: 691-712.

Guveritz, S. 1983. Technology Will Shorten Product Life-Cycles. Business Marketing 12.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. 2009. The WEKA Data Mining Software: An Update; SIGKDD Explorations 11 (1).

Hastie, T. and Tibshirani, R. 1998. Classification by Pairwise Coupling. In: Advances in Neural Information Processing Systems.

Hemetsberger, A. 2001. Fostering cooperation on the internet: social exchange processes in innovative virtual consumer communities. In Proceedings of the Annual ACR Conference. Texas: USA.

Jeppesen, L.B. and Frederiksen, L. 2006. Why do users contribute to firm-hosted user communities? The case of computer-controlled music instruments. Organization Science 17 (1): 45-63.

Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., and Murthy, K.R.K. 2001. Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computation 13 (3): 637-649.

Kira, K. and Rendell, L.A. 1992. A Practical Approach to Feature Selection. In: Ninth International Workshop on Machine Learning, 249-256.

Kononenko, I. 1994. Estimating Attributes: Analysis and Extensions of RELIEF. In: European Conference on Machine Learning, 171-182.

Kozinets, R.V. 1998. On Netnography: Initial Reflections on Consumer Research Investigations of Cyberculture. Advances in Consumer Research 25: 366-371.

Kozinets, R.V. 1999. E-Tribalized Marketing? The Strategic Implication of Virtual Communities of Consumption. European Management Journal 17: 252-264.

Kozinets, R.V. 2006. Click to connect: netnography and tribal advertising. Journal of Advertising Research 46 (3): 279-288.

Kratzer, J. and Lettl, C. 2008. Social Network Perspective of Lead Users and Creativity: An Empirical Study among Children. Creativity and Innovation Management 17 (1): 26-36.

Kratzer, J., Lettl, C., Franke, N. and Gloor, P. A. 2015. The Social Network Position of Lead Users. Journal of Product Innovation Management. doi: 10.1111/jpim.12291

Lakhani, K. 2006. Broadcast search in problem solving: attracting solutions from the periphery. MIT Sloan School of Management, Working Paper.

Latora, V. and Marchiori, M. 2007. A measure of centrality based on network efficiency New J. Phy. 9 (6): 1–16.

Leskovec, J. and Krevl, A. 2014. {SNAP}: A general purpose network analysis and graph mining library in {C++}. [online] Available at: http://snap.stanford.edu/snap [Accessed 01 June 2015].

Lüthje, C. 2000. Kundenorientierung im innovationsprozess. Eine Untersuchung der Kunden-Hersteller-Interaktion in Konsumgütermärkten, Wiesbaden.

Lüthje, C. and Herstatt, C. 2004. The lead user method: an outline of empirical findings and issues for future research. R&D Management 34 (5): 553-568.

Martínez-Torres, M.R. 2014. Analysis of open innovation communities from the perspective of social network analysis. Technology Analysis & Strategic Management 26 (4): 435–51.

Newman, M.E.J. 2010. Networks: An Introduction. Oxford, UK: Oxford University Press.

Olson, E.L. and Bakke, G. 2001. Implementing the lead user method in a high technology firm: a longitudinal study of intentions versus actions. Journal of Product Innovation Management 18 (6): 388-395.

Pajo, S., Verhaegen, P., Vandevenne, D., and Duflou, J.R. 2014. Lead User Identification through Twitter: Case Study for Camera Lens Products. Proceedings of NordDesign 2014. NordDesign, 2014, 294-302.

Pajo, S., Vandevenne D., and Duflou, J.R. 2015. Systematic Online Lead User Identification: Case Study for Electrical Installations. Proceedings of 20th International Conference on Engineering Design, ICED15, June 27-30th, 2015.

Pine, B. J. II 1993. Mass Customization: The New Frontier in Business Competition, Harvard Business School Press, Boston, MA, 1993.

Platt, J. 1998. Fast Training of Support Vector Machines using Sequential Minimal Optimization. Advances in Kernel Methods - Support Vector Learning, MIT Press, Cambridge, MA.

Pollok, P., Lüttgens, D. and Piller, F. 2014. Leading Edge Users and Latent Consumer Needs in Electromobility: Findings from a Nethnographic Study of User Innovation in High-Tech Online Communities. RWTH-TIM Working Paper.

Rajaraman, A. and Ullman, J. D. 2011. Mining of Massive Datasets. Cambridge University Press.

Robnik-Sikonja, M. and Kononenko, I. 1997. An adaptation of Relief for attribute estimation in regression. In: Fourteenth International Conference on Machine Learning, 296-304.

Sabidussi, G. 1966. The centrality index of a graph. Psychometrika 31: 581–603.

Scott, J. 2000. Social Network Analysis: A Handbook 2nd Edition, London, Sage.

Schreier, M. and Prügl, R. 2008. Extending lead-user theory: Antecedents and consequences of consumers' lead userness. Journal of Product Innovation Management 25 (4): 331–346.

Shah, S. 2000. Sources and patterns of innovation in a consumer products field: innovations in sporting equipment. MIT Sloan School of Management, Working Paper 4105.

Spärck Jones, K. 1972. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. Journal of Documentation 28: 11–21.

Sudman, S. 1985. Efficient Screening Methods for the Sampling of Georgraphically Clustered Special Populations. Journal of Marketing Research 22: 20-29.

Twitter. 2015. Twitter Company Info. [online] Available at: https://about.twitter.com/company . [Accessed 15 June 2015].

Twitter Developers. 2015. [online] Available at: https://dev.twitter.com/ [Accessed 29 Jan. 2015].

Von Hippel, E. 1988. The Sources of Innovation. University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship.