INTERNATIONAL CONFERENCE ON ENGINEERING DESIGN, ICED15
27-30 JULY 2015, POLITECNICO DI MILANO, ITALY

SYSTEMATIC ONLINE LEAD USER IDENTIFICATION - CASE STUDY FOR ELECTRICAL INSTALLATIONS

Pajo, Sanjin; Vandevenne, Dennis; Duflou, Joost R.
KU Leuven, Belgium

Abstract

Identification of emerging needs and partial solutions is crucial for industries to stay competitive in a fast evolving marketplace. A small subgroup of customers, called lead users have been shown to experience needs before the rest of the marketplace and propensity to find solutions to address those needs. Involvement of lead users in the new product development process leads to attractive and successful new products. In this paper, the authors present a fast lead user identification approach that makes use of data mining, network analysis and machine learning techniques. An implementation of the approach for the micro-blogging site Twitter is described and the results of the effectiveness analysis case in the domain of electrical installations are reported. The implemented methodology points to opportunities in systematic and fast identification of lead users online with additional studies for different domains required to validate the approach.

Keywords: Lead User Identification, Crowdsourcing and Funding, Open Innovation, Data Mining, Social Media

Contact:
Sanjin Pajo
KU Leuven
Mechanical Engineering
Belgium
sanjin.pajo@kuleuven.be

Please cite this paper as:
Surnames, Initials: *Title of paper*. In: Proceedings of the 20th International Conference on Engineering Design (ICED15), Vol. nn: Title of Volume, Milan, Italy, 27.-30.07.2015

1 INTRODUCTION

Innovation is an important competitive weapon and driver of the economy. A significant portion of innovation is carried out by a subgroup of customers, called lead users (Von Hippel, 1986). They are consumers that "face needs that will be general in a marketplace – but face them […] before the bulk of the marketplace encounters them, and […] are positioned to benefit significantly by obtaining a solution to those needs " (Von Hippel, 1986). Lead users not only have propensity to engage in innovation, but also help develop commercially attractive products and offer solutions to emerging challenges with existing product use (Schreier et al., 2007). An example given by Shah (2000) describes windsurfers, who were developing new aerial tricks and, having difficulties remaining on the board, realized the need for foot straps which were later made commercially available. Lead users also exhibit considerably higher levels of innovation than ordinary users (Schreier & Prügl, 2008; Olson & Bakke, 2001). Schuhmacher and Kuester (2012) state that companies should engage lead users in the design process and should focus on specific lead user characteristics to find those individuals that can contribute to more effective product development.

Today's consumers are using online media and web applications to communicate with other consumers, to share product experiences and innovation ideas. There is also a rapid development of social networking sites, i.e. Twitter, Facebook, Wikipedia and YouTube, platforms that support the exchange of not only product knowledge and use experiences, but also of emerging needs and solutions. Openness, a trait of social networks and willingness to entertain new technological and social trends have a positive impact on consumer's product knowledge and innovativeness (Füller et al., 2008). There are less communication barriers and users are exposed to a variety of different sources and often unfiltered information (Assael, 1998; Kozinets, 2002; Pitta & Fowler, 2005). A clear probability of identifying users with exceptionally strong lead user characteristics exists in large online communities like social networks (Bilgram et al., 2008). The focus of the paper is exploration of automated classification of lead users and non-lead users through social media by making use of data mining and machine learning techniques.

2 RELATED RESEARCH

Currently, the most utilized method for identifying lead users is the so-called Lead User Method (LUM). It makes use of questionnaires and interviews performed by experts to identify relevant users. The method consists of different approaches, including screening, pyramiding and broadcasting. Screening selects individuals from a sample of the entire population to be evaluated for lead user qualifications. Low efficiency and heavy reliance on interviews performed by experts make the approach costly in terms of time and human resources. Pyramiding starts by evaluating a small number of individuals and based on recommendations works its way up the pyramid of expertise to lead users. It is based on the notion that experts in a particular domain will know individuals who tend to know more or are more engaged in the same domain. Pyramiding is more efficient than screening, requiring 30% less effort (Von Hippel et al., 2008). Researchers can 'move up the pyramid' to find those users with more lead user attributes in a particular domain (Von Hippel et al., 2008). Efficiency of the pyramiding approach highly depends on the relationship between the individuals in the target population, their knowledge and familiarity with one another (Von Hippel et al., 2008). The approach minimizes in efficiency with the lack of observability of other individuals within the sample and lack of need to share or advertise (Von Hippel et al., 2008). In the broadcasting approach, a problem is sent to a group of potential solvers outside a company (Lakhani, 2006; Hienerth et al., 2007). The cost of identifying a group of potential problem solvers is not clearly defined and it is time and resource consuming with candidates selected that may not exemplify lead user characteristics. LUM approaches are often susceptible to dishonest replies and systematic bias. Additionally, high resource and time costs, since the identification process can last more than six months, and reliance on self-assessment of respondents during the initial contact phase, reduce the efficacy and efficiency of the method. For example, in a study of the LUM, as described by Olson and Bakke (2001), managers discovered a number of obstacles in executing the approach; "the efforts required to find, qualify, and recruit experts for trend analysis and lead users for concept generation were seen as the most burdensome tasks." Olson and Bakke (2001) were stopped short of having the lead users method utilized as a part of the product development team in a high tech firm because of the extra time required to find experts for trend analysis and lead users for concept generation.

Netnography (Kozinets, 1999, 2002) is another approach aimed at identifying lead users with online communities as the target platform. Prior to the identification process, user characteristics are determined and a target community is explored and selected. The initial part of the method consists of observations of online community behaviour. The following steps involve qualitative analysis of member communications and discussions leading to interviews with identified potential lead users (Belz & Baumbach, 2010). In a Netnography study of basketball communities, researchers observed five communities over a period of six months and analyzed more than 240,000 online posts (Füller et al., 2007). Similarly to the Lead User Method, the Netnography approach is resource and time consuming, requiring experts to analyze large number of user online discussions. Still, Netnography studies exemplify great opportunities in systematic identification of lead users online.

In this paper, the approach proposed aims to reduce the significant time and resource costs associated with the aforementioned lead customer identification methods by using a fast and systematic approach utilizing data mining and machine learning algorithms. In the next section, the methodology of the semi-automated approach for a micro-blogging site is presented.

3 SYSTEM ARCHITECTURE

The proposed systematic approach, referred to as Fast Lead User IDentification (FLUID), consists of gathering of consumer product related online activity and posts, which are computationally treated. Currently, the approach is being tested on the social networking site Twitter. Twitter was selected because it has a large user base, over 300 million, with much of the discussion and user metadata publicly accessible through its Application Programming Interface (API) (Dev.twitter.com, 2014). Twitter is also a platform where posts or tweets regarding product use and development are commonplace (Pak & Paroubek, 2010) and users can be contacted through Twitter accounts for further evaluation or engagement in design process. Figure 1 below presents the semi-automated FLUID approach implemented for Twitter that is described in the following sections with each component of the process shown, including query retrieval, data gathering and consolidation, data processing and data evaluation. Search keyword generation, the only non-automated part of the approach, and data collection are described in the next section.

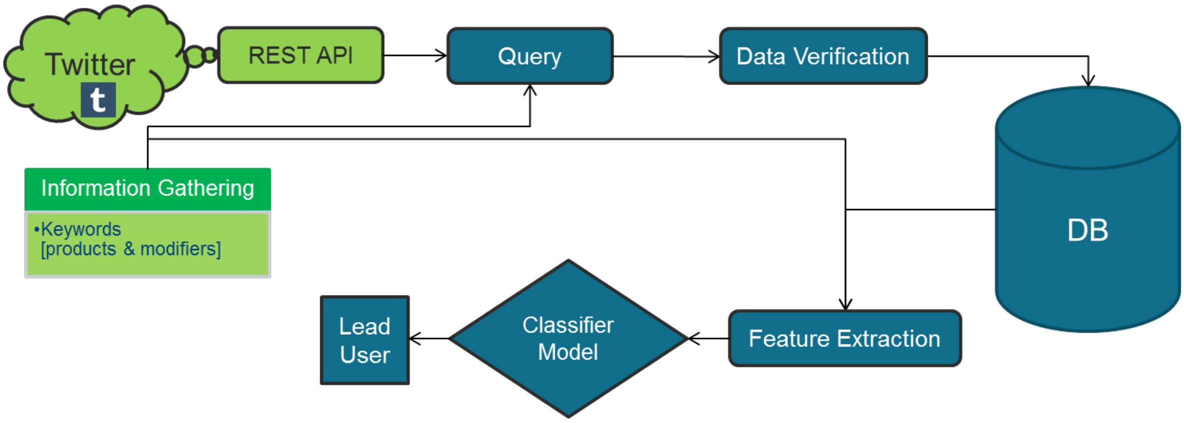

Figure 1. Fast Lead User Identification through Twitter

3.1 Keyword Generation and Data Collection

To begin collecting Twitter data, a set of relevant keywords is gathered as input for the Twitter user search process. Currently, for each validation case, the stakeholder or the company is asked to select a set of word bigrams consisting of a product name or domain (i.e. switch, domotics) and preferably a parameter or a modifier (i.e. wireless, build). The keywords are selected in accordance with the business value they have for the company. Each keyword set is given a significance value, an estimate of its importance to the stakeholders. A few examples of collected keywords for electrical installations are shown in Table 1 below:

Table 1. Example of search keywords for lighting control products

Value	Term A	Term B
39	electrical	installation
31	energy	saving
25	assisted	living
14	innovative	control

After search keyword selection, the next step is retrieval and storage of Twitter data for the classification process. For this purpose, a test platform was built to perform data retrieval, storage, feature extraction and classification. The test platform makes use of the Twitter search engine and sets of collected bigrams to retrieve relevant tweets and accompanying user metadata. The collected data is verified for accuracy and redundancy and thereafter stored inside a database, as shown in Figure 1. Additionally, for each user, the list of user's Twitter followers and friends is retrieved for later network construction and analysis. The data collection process has an indeterminate length but may be time limited depending on the stakeholder's needs. The next step in the process is feature extraction for each retrieved user and finally, classification of Twitter users into a lead user or a non-lead user.

3.2 Feature Extraction and Classification

After data collection is complete, metadata are analysed per user and a set of features is compiled for each user retrieved from Twitter. *Frequencies*, *relevance*, *sentiment*, *trend* and *network* measures are extracted from the collected metadata for each Twitter user. *Frequencies* are a representation of user's behaviour on the social networks, for example the rate at which the user is tweeting and adding followers, friends, favourites. As shown in existing research, customer behaviour and engagement are factors that delineate a lead user from a non-lead user (Bilgram et al. 2008).

A second measure is *relevance*, which is a numerical estimate of the user's timeline (all tweets) relevance to the set of search keywords collected by the stakeholder. The test platform makes use of term frequency-inverse document (tweet) frequency numerical statistics (Wikipedia, 2014) to calculate the relevance score for each user. The score is computed as follows:

$$score(Q_T, c) = \sum_{(t)\epsilon Q_T,c} tf(t,c) \cdot \ln\frac{|C|}{df(t,C)} \tag{1}$$

where Q_T is the collection of keywords, $|C|$ is the size of the tweet collection C, $tf(t,c)$ is the number of times the keyword t appears in the tweet c and $df(t,C)$ is the number of tweets in C that contain t. Thereafter, the relevance score of the tweet collection or the relevance score for the Twitter user N is computed by:

$$score(Q_T, C_N) = \frac{\sum_{c\in C_N}(score(Q_T,c))}{|C_N|} \tag{2}$$

Relevance is closely related to use experience and knowledge of the product or the specific domain, two essential characteristics in lead user identification (Bilgram et al. 2008).

Dissatisfaction with existing solutions in the marketplace is another validated (Bilgram et al. 2008; Schuhmacher and Kuester, 2012) lead user characteristic and through the test platform emotional disposition of user's tweets or *sentiment* is calculated by making use of the SentiWordNet lexicons (Sentiwordnet.isti.cnr.it, 2014) to aggregate the percent of positive, negative and neutral tweets. The fourth measure, *trend* is a numerical indication of how early in a user's timeline relevant keyword hashtags are used. As stated previously, for each user, lists of Twitter followers and friends are retrieved, which allows a graphical representation of collected networks, with Twitter users as nodes and relationships (friendships, followers) as edges. For each user, *network* centrality measures are derived and added to the feature set. For example, Kratzer and Lettl (2008) have shown that higher measures of centrality have positive correlation characteristics of being a lead user. Online users not only absorb knowledge, but also share the knowledge with the online communities exhibiting opinion leadership qualities (Jeppesen & Laursen, 2009). Lead users foster information sharing and help diffuse new needs and solutions in their networks (Bilgram et al., 2008; Kratzer and Lettl, 2008). The extracted feature scores are standardized, normalized and tabulated to be used as an input as test data for the classification model.

The classification model is created using a known set of Twitter lead and non-lead users, collected by making use of a self-administered online survey that measured five validated characteristics of being a lead user: dissatisfaction, ahead of trend, product knowledge, use experience and opinion leadership (Pajo et al., 2014a). Cross-validation was performed on the training set and the random tree algorithm (Wiki.pentaho.com, 2014) was selected to generate the classification model due to its overall high accuracy. The WEKA platform (Hall et al., 2009) generated tree representation of the model is shown in the Figure 2 below, which is an improvement on the tree model view reported by Pajo et al., (2014b) due to a larger set of features and corrections in metadata collections.

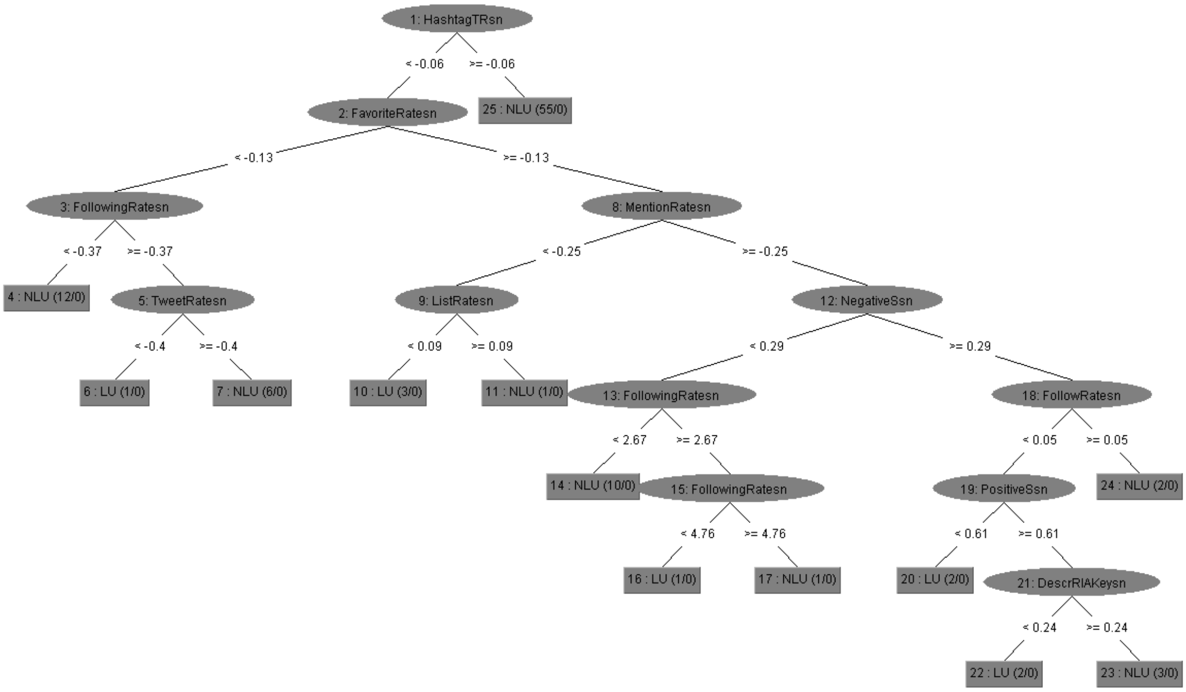

Figure 2. Tree view of the classification model

Item	Meaning
HashtagTRsn	Hashtag ahead of trend score
FavoriteRatesn	Rate of favorites added per day
FollowingRatesn	Rate of users followed per day
MentionRatesn	Rate of Twitter user mentions per tweet
TweetRatesn	Rate of tweets per day
ListRatesn	Rate of lists per day
NegativeSsn	Percent of tweets with negative disposition
FollowRatesn	Rate of followers added per day
PositiveSsn	Percent of tweets with positive disposition
DescriptionRIAKeysn	Relevance of the user's description to the keywords
LU	Lead User
NLU	Non-Lead User

Figure 2 shows how extracted user features play a role in how the model delineates between a lead user (LU) and a non-lead user (NLU). Trend score (HashtagTRsn), frequencies (FavoritesRatesn, MentionRatesn, TweetRatesn, ListRatesn, FollowingRatesn, FollowRatesn), emotional sentiment (PositiveSsn, NegativeSsn) and relevance (DescriptionRIAKeysn) normalized scores allow for separating a lead user from a non-lead user. The network centrality measures do not feature in model tree view, which depends greatly on the set of collected training data. In the next section, a case study is described for the FLUID approach.

4 EFFECTIVENESS ANALYSIS

A study was performed with an industrial partner to evaluate the feasibility and the effectiveness of the approach through Twitter in the domain of electrical installations. At first, the in-company R&D team, was asked to provide a set of search keywords. The keywords were used as input to retrieve relevant tweets and associated user metadata. The automated data collection process lasted four weeks due to the Twitter API limitation restrictions, 180 calls every 15 minutes (Dev.twitter.com, 2014). Additionally, data older than one week could not be retrieved (Dev.twitter.com, 2014). In the first two weeks the test platform collected data using the provided keywords in an automated loop mode and during the second week friends and follower lists were obtained for each Twitter user. After the classification, two in-company assessors were asked to evaluate a list of 67 classified users by indicating whether the Twitter user would be of interest for future collaboration. For each user, descriptive metadata and a set of most relevant tweets to the set of keywords was presented to the company assessors. The list of 67 Twitter users included 34 predicted lead users and 33 predicted non-lead users uniformly selected from the classification set. Prior to evaluation, the assessors were informed that the list of 67 Twitter users contained both lead and no-lead users but not of the exact numbers for each class. The order of Twitter users was changed for each reviewer to minimize the ordering bias. In the next section, the results of the evaluations are described.

4.1 Results

After a discussion, in-company assessors agreed together that 5 of 34 predicted lead users would be most relevant to contact for further collaboration, but expressed doubt in selected users being lead users for the company. For the remaining predicted 29 lead users, it was difficult to know if they are profiles created by competitors and the assessors chose not to select the cases.

Table 2. Overall results of the two evaluations

		Company Assessors	
		NLU	**LU**
FLUID platform	**NLU**	33	0
	LU	29	5

Results are shown in Table 2 and give the overall accuracy for the classification model of 56.71%. The precision for lead user class is a very low, 15%, and for the non-lead user class 100%. The recall is 100%. The effectiveness analysis results show a very low agreement between the company assessors and the test platform with respect to the classified lead users and a high agreement with respect to classified non-lead users. The reviewers were not fully convinced that these Twitter cases could be their lead users and although they expressed interest in proceeding to the next step of contacting selected Twitter users, additional interviews would be necessary for validation. Unlike the previous study on camera lens products (Pajo et al., 2013), that indicated that the model can be used in separating online lead users from non-lead users, the results for the domain of electrical installations are inconclusive.

After the evaluation stage, the assessors stated that the selected keywords were not optimal. For example, the word *switch* is in most cases used in a completely different manner than the meaning that we had given to it (switch = button, control). Selection of product keywords can cause ordering and selection biases. Semantic review of the keywords is an additional step that is needed where certain synonyms would have given better results. The assessors expressed desire to adapt their keywords for follow up iterations.

Results also do not reflect the high level of expressed disagreement in selection of lead users between the reviewers for the two iterations. Even with expert analysis, the concept of a lead user differs from person to person. The assessors suggested that a persona or representation of targeted lead user is made to aid the evaluation process. Additionally, tweets are too short and vague and often times not related to the target topic, which makes it difficult to assess the profile. For a more effective analysis, additional metadata and analytics are included for the follow up cases.

Significant doubt exists as to the effectiveness of the methodology. Considering the feedback from the industrial partner, additional improvements are envisioned, including improvements in the classification model and gathering of search keywords. With regards to time and resource utilization, the FLUID method offers a significantly faster approach. The FLUID identification through Twitter currently takes around 4 weeks as opposed to over 6 months for the LUM, but the process can be considerably sped up by Twitter data purchase, where data collection is not subject to Twitter API restrictions. With data purchase, relevant tweet data older than one week would also be accessible through the Twitter search engine. The significantly reduced duration of the identification process makes the approach also suitable for manufacturers where the product cycle is short. The methodology can be extended to other web frameworks and social media sites, with the information retrieval techniques tailored for the available user and network metadata. Overall, the findings point to opportunities in lead user identification through social media and further extend the web based lead user methods like Netnography.

5 CONCLUSION

This paper contributes to the fuzzy-end front end of product development field by applying the lead user theory on a social networking micro-blogging site, Twitter in an industrial context. It is potentially time and resource saving methodology to lead user identification, where data mining and machine learning techniques rapidly extract and classify the available user online discussions and metadata but additional validation is necessary to verify its effectiveness. The empirical results for the explored cases are significant for the practitioners in design and innovation management as the studies signal opportunities for the private sector in identifying emerging needs, trends and partial solutions through social media sites. Additional industrial case studies through a variety of social networking sites are expected with the aim to validate the FLUID approach.

REFERENCES

Assael, H. (1998) Consumer Behavior and Marketing Action. Cincinnati: South Western College Publishing.

Belz, F.M. and Baumbach, W. (2010) Netnography as a Method of Lead User Identification. Creativity and Innovation Management, Vol. 19, No. 3, pp. 304-313.

Bilgram, V., Brem, A. and Voigt, K.I. (2008) User-Centric Innovations in New Product Development - Systematic Identification of Lead Users Harnessing Interactive and Collaborative Online Tools. International Journal of Innovation Management, Vol. 12, No. 3, pp. 419-458.

Dev.twitter.com, (2014) Twitter Developers. [online] Available at: https://dev.twitter.com/ [Accessed 10 Dec. 2014].

Füller, J., Jawecki, G. and Muhlbacher, H. (2007) Innovation Creation by Online Basketball Communities. Journal of Business Research, Vol. 60, No. 1, pp. 60-71.

Füller, J., Matzler, K. and Hoppe, M. (2008) Brand Communitiy Members as a Source of Innovation. Journal of Product Innovation Management, Vol. 25, No. 6, pp. 608-619.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009) The WEKA Data Mining Software: An Update; SIGKDD Explorations, Vol. 11, No 1.

Hienerth, C., Poetz, M. and Von Hippel, E. (2007) Exploring key characteristics of lead user workshop participants: Who contributes best to the generation of truly novel solutions? DRUID Summer Conference 2007, Copenhagen, June 18-20.

Jeppesen, L.B. and Laursen, K. (2009) The role of lead users in knowledge sharing. Research Policy. Vol. 38, No. 10, pp. 1582-1589.

Kozinets, R. (1999) E-Tribalized Marketing? The Strategic Implications of Virtual Communities of Consumption. European Management Journal, Vol. 17, No. 3, pp. 252-264.

Kozinets, R. (2002) The Field Behind the Screen: Using Netnography for Marketing Research in Online Communities. Journal of Marketing Research, Vol. 39, No. 1, pp. 61-72.

Kratzer, J. and Lettl, C. (2008) A Social Network Perspective of Lead Users and Creativity: An Empirical Study among Children. Creativity and Innovation Management, Vol. 17, No. 1, pp. 26-36.

Lakhani, K. (2006) Broadcast search in problem solving: attracting solutions from the periphery. Technology Management for the Global Future, Vol. 4, No. 1, pp. 2450-2468.

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, Vol. 18, No. 6, pp. 388-395.

Pajo, S., Verhaegen, P., Vandevenne D. and Duflou, J. (2013) Analysis of Automatic Online Lead User Identification. Proceedings of the 23rd CIRP Design Conference, Bochum, Germany, March 11th, pp. 505-514.

Pajo, S., Verhaegen, P., Vandevenne, D. and Duflou, J. (2014a) Lead User Identification through Twitter: Case Study for Camera Lens Products. Proceedings of NordDesign 2014 : 10th biannual conference on design and development. NordDesign Conference, 2014.

Pajo, S., Verhaegen, P., Vandevenne, D. and Duflou, J. (2014b) Fast Lead User Identification Framework. TRIZ Future Conference 2014. Lausanne, 29-31, October 2014.

Pak, A. and Paroubek, P. (2010) Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Proceedings of the Seventh International Conference on Language Resources and Evaluation, pp. 1320-1326.

Pitta, D. A. and Fowler D. (2005) Internet community forums: an untapped resource for consumer marketers. Journal of Consumer Marketing, Vol. 22, No. 5, pp. 265-274.

Schreier, M., Oberhauser, S. and Prügl, R. (2007) Lead users and the adoption and diffusion of new products: Insights from two extreme sports communities. Marketing Letters, Vol. 18, No. 1-2, pp. 15-30.

Schreier, M. and Prügl, R. (2008) Extending Lead User Theory: Antecedents and Consequences of Consumer's Lead Userness. Journal of Product Innovation Management, Vol. 25, No. 4, pp. 331-46.

Schuhmacher, M. and Kuester, S. (2012) Identification of Lead User Characteristics Driving the Quality of Service Innovation Ideas. Creativity and Innovation Management, Vol. 21, No. 4, pp. 427-442.

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.

Sentiwordnet.isti.cnr.it, (2014). SentiWordNet. [online] Available at: http://sentiwordnet.isti.cnr.it/ [Accessed 10 Dec. 2014].

Von Hippel, E. (1986) Lead Users: A Source of Novel Product Concepts, Management Science, Vol. 32, No. 7, pp. 791-806.

Von Hippel, E., Franke, N. and Prügl, R. (2008) 'Pyramiding': Efficient Identification of Rare Subjects. MIT Sloan School of Management, Working paper, pp. 4719-4708.

Von Hippel, E., Thomke, S. and Sonnack, M. (1999) Creating Breakthroughs at 3M. Harvard Business Review Vol. 77, No. 5, pp. 47-57.

Wiki.pentaho.com, (2014). RandomTree - Pentaho Data Mining - Pentaho Wiki. [online] Available at: http://wiki.pentaho.com/display/DATAMINING/RandomTree [Accessed 10 Dec. 2014].

Wikipedia, (2014). Tf–idf. [online] Available at: http://en.wikipedia.org/wiki/Tf%E2%80%93idf [Accessed 15 Dec. 2014].